Neurocomputing 182 (2016) 284–293

Contents lists available at ScienceDirect

Neurocomputing

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

Pairwise clustering based on the mutual-information criterion

Amir Alush, Avishay Friedman, Jacob Goldberger *

Faculty of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel

ARTICLE INFO

Article history:
Received 12 June 2015
Received in revised form
3 December 2015
Accepted 3 December 2015
Communicated by X. Gao
Available online 23 December 2015

Keywords:
Graph clustering
Pairwise clustering
Mutual information
Spectral clustering
Normalized-cut

ABSTRACT

Pairwise clustering methods partition a dataset using pairwise similarity between data-points. The pairwise similarity matrix can be used to define a Markov random walk on the data points. This view forms a probabilistic interpretation of spectral clustering methods. We utilize this probabilistic model to define a novel clustering cost function that is based on maximizing the mutual information between consecutively visited clusters of states of the Markov chain defined by the similarity matrix. This cost function can be viewed as an extension of the information-bottleneck principle to the case of pairwise clustering. We show that the complexity of a sequential clustering implementation of the suggested cost function is linear in the dataset size on sparse graphs. The improved performance and the reduced computational complexity of the proposed algorithm are demonstrated on several standard datasets and on image segmentation task.

1. Introduction

Effective automatic grouping of objects into clusters is one of the fundamental problems in machine learning and in other fields of study. In many approaches, the first step toward clustering a dataset is extracting a feature vector from each object. This reduces the problem to the aggregation of groups of vectors in a feature space. A commonly used algorithm in this case is the *k*-means. One drawback of *k*-means is that it can only find clusters that are linearly separable in the feature space. Furthermore, in many cases features representation is not available and we are only given pairwise similarity information between data points. For example, in social networks, only binary neighborhood relations are given. In these cases feature based clustering algorithms cannot be applied in a straightforward way. Instead, we seek for a partition of the data based only on the similarity measure between the points.

The problem of pairwise clustering can be naturally viewed as a graph clustering where the data points are associated with the graph nodes and the pairwise affinities are the weights on the edges. We want to find a partition of the graph such that the edges between different groups have low weights and the edges within a group have high weights. Out of the numerous pairwise clustering algorithms, spectral clustering has gained considerable attention in recent years due to its strong performance on arbitrary shaped clusters, and its well-defined mathematical framework. Spectral clustering algorithms [1–5] are based on finding a low dimensional embedding using eigenvector computation which can be slow. The Power Iteration Clustering (PIC) [6] is a variant of spectral clustering that directly finds the low-dimensional embedding. Graclus [7] is another efficient graph clustering algorithm that is based on directly optimize the Ncut score using multilevel kernel *k*-means and avoids the eigenvector computations.

Another family of clustering algorithms, that are derived from information-theory concepts, corresponds to the case of distributional clustering. Here each data point is described as a distribution. This situation is illustrated by the generic example of document clustering based on word histograms [8,9]. In this case, the mutual information (MI) between word occurrences and clusters of documents is a natural clustering criterion that has been proven to be powerful in many cases [10,11]. Given a clustering task, we look for a clustering that maximizes the mutual information between cluster labels and features of data points. In other words, we search for a clustering that minimizes the information loss in the feature space caused by shifting from points to clusters. Information-theoretic approaches have been intensively used for data clustering algorithms (see e.g. [12–15]). The information-theoretical principle described above, however, is applicable when a feature distribution, associated with each data point, is provided as part of the problem setup. It is not straight-forward how to adapt this information theoretic principle for the problem of graph clustering where only pairwise similarity is given.

In this paper we extend the mutual information clustering criterion to the domain of pairwise clustering. The probabilistic

* Corresponding author.
E-mail addresses: amiralush@gmail.com (A. Alush), favishay@gmail.com (A. Friedman), jacob.goldberger@biu.ac.il (J. Goldberger).

http://dx.doi.org/10.1016/j.neucom.2015.12.025
0925-2312/

interpretation of spectral clustering, based on a Markov random walk, is used to associate a distribution with each data point via the corresponding conditional distribution row in the Markov transition matrix. In particular, we define a random walk on the data points and maximize the mutual information between cluster labels of data-points that are visited during the random walk. We show that this results in a clustering cost function, alternative to the Normalized-Cut criterion, that yields improved performance clustering on real-world datasets.

In this study we apply an information theoretic framework to pairwise clustering as an alternative to spectral clustering. There are other methods that combine information theory with spectral clustering. Jenssen et al. [16] used an information-theoretic distance measure to define a graph-cut criterion for clustering. Another approach is based on estimation of Renyi's entropy to define optimal clustering in terms of certain spectral properties of the affinity matrix [17]. Unlike these methods, we directly address the pairwise matrix (with no need for feature vectors) and we use entropy to analyze the random walk along the clustered data points instead of measuring intra-variability of the data points.

The remainder of this paper is organized as follows. Section 2 defines the notation of similarity graphs and random walk on the graph nodes. Section 3 describes the minimum information-loss criterion for clustering the Markovian random-walk states. Section 4 introduces the Information-Theoretic Pairwise Clustering (ITPC) algorithm. Section 5 describes the relation of the proposed algorithm to the other information-theory based algorithms. Section 6 describes numerical experiments on several standard datasets and Section 7 presents comparative results on image segmentation dataset. A preliminary version of this work was presented at the SIMBAD Workshop, York, England, 2013.

2. Similarity graphs and random walks

Given a set of data points $x_1,\ldots,x_n$ and some symmetric notion of similarity $w_{ij}\geq 0$ between all pairs of data points x_i and x_j, the goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. In the common case where the data points live in the Euclidean space R^d, a reasonable candidate for a similarity measure is the Gaussian function $w_{ij}=\exp(-\|x_i-x_j\|^2/(2\sigma^2))$ (where the parameter σ controls the width of the local neighborhoods). Ultimately, the choice of the similarity function depends on the domain the data come from and the specific clustering task. In a more general case we do not have an explicit representation for each data point by a feature vector. Instead, the only available information for data clustering is pairwise similarities between the data points.

We can represent the dataset of n points and pairwise similarities $\{w_{ij}\}$ as a similarity graph $G=(V,E)$. Each vertex in this graph represents a data point. Two vertices $i,j\in V$ are connected if the similarity w_{ij} between the corresponding data points is positive and the edge is weighted by w_{ij}. The problem of clustering can now be reformulated using the similarity graph: we want to find a partition of the graph in which existing edges between different groups have low weights and edges within a group have high weights. The normalized-cut score is one of the popular ways to translates this intuition into a formal clustering criterion.

Denote the similarity weight matrix by $W=(w_{ij})$. For two (not necessarily disjoint sets) $A,B\subset V$ we define

$$W(A,B)=\sum_{i\in A,j\in B} w_{ij}. \tag{1}$$

The degree of a vertex $i\in V$ is defined as

$$d_i=\sum_{j=1}^{n} w_{ij}=W(\{i\},V).$$

The volume of $A\subset V$ is

$$\mathrm{vol}(A)=\sum_{i\in A} d_i=W(A,V). \tag{2}$$

The normalized-cut score [2,18] of a given partitioning of the graph nodes into m disjoint subsets $\{A_1,\ldots,A_m\}$ is

$$\mathrm{Ncut}(A_1,\ldots,A_m)=\sum_{i=1}^{m}\frac{W(A_i,\overline{A}_i)}{\mathrm{vol}(A_i)} \tag{3}$$

such that $\overline{A}_i$ is the complement set of A. In the clustering that minimizes this score, edges between different groups have low weights. The role of dividing by $\mathrm{vol}(A_i)$ is to ensure that the cluster sizes (as measured by edge weights) are balanced. Minimizing the Ncut score, however is NP hard even for $m=2$ [19]. The Ncut spectral clustering algorithm [2,18] is an algorithm that finds an optimal solution for a relaxation of the Ncut criterion (3). All variants of the spectral clustering algorithm are based on using eigenvectors of the Laplacian matrix of the similarity graph to represent the abstract data points as points in the Euclidean space. The clusters can be then obtained by applying simple clustering algorithms such as k-means in the embedded space [1–3]. Dhillon et el. [7] applied kernel k-means to directly optimize the Ncut score.

The Ncut score is defined using a graph theory formulation (3). Meila and Shi [20] provided a probabilistic interpretation of it as a criterion for clustering the states of the random walk defined by the similarity matrix W. Define the degree matrix D as the diagonal matrix with the degrees $d_1,\ldots,d_n$ on the diagonal. The $n\times n$ matrix $P=D^{-1}W$ is a stochastic matrix (non-negative entries, row sums are all 1). Using the transition matrix P we can define a stationary Markov chain that corresponds to a random walk on the graph nodes. Let $X=\{X_t\}$ be the n-valued stationary Markov chain defined by

$$P_{ij}=(D^{-1}W)_{ij}=p(X_2=j|X_1=i)=\frac{w_{ij}}{d_i} \tag{4}$$

The transition probability P_{ij} of jumping in one step from i to j is proportional to the edge weight w_{ij}. Let $\pi=(\pi_1,\ldots,\pi_n)$, where $\pi_i=d_i/(\sum_j d_j)$. It can be easily verified that $P^{\top}\pi=\pi$. Hence, if the graph is connected and non-bipartite, then π is the unique stationary distribution of the Markov chain defined by P [5]. Therefore, the joint stationary probability of X_1 and X_2 is

$$p(X_1=i,X_2=j)=\frac{w_{ij}}{\mathrm{vol}(V)}. \tag{5}$$

Given the random walk model (4) we can translate the pairwise clustering problem, into the problem of clustering the states of a Markov chain. Let A and B be two subsets of V. From Eq. (5) we obtain that

$$p(X_2\in B|X_1\in A)=\frac{W(A,B)}{\mathrm{vol}(A)}. \tag{6}$$

Let $\{A_1,\ldots,A_m\}$ be a partition of the n graph nodes into m clusters. Substituting Eqs. (6) in (3), we obtain the following probabilistic interpretation of the Ncut score [20]:

$$\mathrm{Ncut}(A_1,\ldots,A_m)=\sum_{i=1}^{m} p(X_2\notin A_i|X_1\in A_i). \tag{7}$$

This interpretation of Ncut tells us that when minimizing Ncut, we actually look for a graph partition such that a random walk seldom transitions from one cluster to another.