Neurocomputing 90 (2012) 72–84

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Approximation techniques for clustering dissimilarity data ☆

Xibin Zhu, Andrej Gisbrecht, Frank-Michael Schleif, Barbara Hammer *

Bielefeld University, CITEC Centre of Excellence, D-33594 Bielefeld, Germany

ARTICLE INFO

Available online 21 March 2012

Keywords:
Clustering dissimilarity data
Approximation
Patch processing
Nyström approximation
Affinity propagation
Generative topographic mapping
Neural gas

ABSTRACT

Recently, diverse high quality prototype-based clustering techniques have been developed which can directly deal with data sets given by general pairwise dissimilarities rather than standard Euclidean vectors. Examples include affinity propagation, relational neural gas, or relational generative topographic mapping. Corresponding to the size of the dissimilarity matrix, these techniques scale quadratically with the size of the training set, such that training becomes prohibitive for large data volumes. In this contribution, we investigate two different linear time approximation techniques, patch processing and the Nyström approximation. We apply these approximations to several representative clustering techniques for dissimilarities, where possible, and compare the results for diverse data sets.

1. Introduction

The amount of digital data doubles roughly every 20 months. Hence automatic tools to deal with large data sets become indispensable for humans to extract relevant information from the data. In this context, clustering constitutes one of the standard techniques to structure and compress large data sets. Algorithms which represent clusters in an intuitive form such as prototype-based techniques offer the possibility to directly inspect the results. Additional functionality such as e.g. topographic mapping as enabled by self-organizing algorithms can provide further inside into the data structure. Because of these facts, prototype based clustering and extensions towards topographic mapping have lost none of its attraction for users from diverse application areas [19].

Not only the size of modern data sets, but also its complexity increases rapidly in modern application areas. Improved sensor technology, for example, leads to very high dimensional measurements corresponding to a very detailed resolution of the available information. At the same time, dedicated data formats such as XML files, network data, graph structures and the like become more and more common. Classical prototype based classification such as *k*-means clustering, neural gas, or the self-organizing map usually deals with Euclidean vectors only. Hence these algorithms are no longer suited in these settings. While the Euclidean distance yields to almost meaningless values for high dimensionality, a lossless vectorial representation is not even possible for data structures such as sequences, trees, or graph structures.

This fact has led to a variety of extensions of prototype-based techniques to deal with more complex data formats, see e.g. [3]. One prominent interface is offered by a general similarity or dissimilarity matrix: only pairwise similarities or dissimilarities of data have to be defined based on which learning takes place. Various dissimilarity measures are available for dedicated data formats: for example, alignment for sequences [14], functional norms for functional data [26], divergences for probability distributions [27], graph and tree kernels [16], or the compression distance for general symbolic sequences [7]. Hence a formulation in terms of dissimilarities extends the applicability of prototype based techniques to a large variety of modern application areas. Since data are characterized by pairwise relations rather than Euclidean vectors, we refer to 'relational data' in the following.

There exist different principled ways to transfer prototype based clustering towards dissimilarity data: kernel methods extend standard techniques towards more general data by means of kernelization, see e.g. [29,6]. This has the drawback that a valid kernel has to be present, i.e. data have to be inherently Euclidean corresponding to a positive semidefinite Gram matrix. Techniques such as proposed in [17] are based on the so-called dual cost function associated to quantization error. It allows an elegant solution using techniques of statistical physics. However, no explicit prototypes are available in this setting. As an alternative, exemplar based techniques restrict prototype positions to data points, such that standard cost functions as the quantization error are still defined for general dissimilarities. Optimization of these costs, however, becomes hard due to the discrete space of feasible solutions. Median clustering determines optima by extensive search [20,8]. It often leads to only suboptimal solutions due to

☆This work has been supported by the DFG under Grant number HA2719/4-1 and by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative.
* Corresponding author. Tel.: +49 521 106 12115; fax: +49 521 106 12181.
E-mail addresses: xzhu@techfak.uni-bielefeld.de (X. Zhu), agisbrecht@techfak.uni-bielefeld.de (A. Gisbrecht), fschleif@techfak.uni-bielefeld.de (F.-M. Schleif), bhammer@techfak.uni-bielefeld.de (B. Hammer).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.01.033

the search in a very restricted space [8]. Recently, a very promising alternative optimization method has been proposed [10]: affinity propagation (AP) reformulates the quantization error such that it can be formalized as a factor graph, for which powerful optimization techniques such as the max-sum algorithm are readily available. By relying on log-likelihood values, the algorithm inherently deals with a continuous relaxation of the discrete optimization problem and usually arrives at very good optima of the cost function. We will consider AP as one very promising clustering technique for exemplar based data representation for general similarities in the following.

AP has the drawback that additional functionality such as neighborhood preservation of the clusters is not available. Recently, Euclidean prototype based topographic mapping as offered by neural gas [23] and the generative topographic mapping [4] have been extended to general dissimilarities. A key technique is an implicit embedding of data in pseudo-Euclidean space [15,13]. This way, a continuous update of prototypes becomes possible leading to robust solutions of the respective cost function in pseudo-Euclidean space. Using a simple algebraic equation, the explicit computation of the embedding becomes superfluous—update rules which depend on the given dissimilarities only can be derived. We will consider two instantiations of this technique, relational neural gas (RNG) and relational generative topographic mapping (RGTM). These techniques constitute two important schemes to infer a topographic mapping for general dissimilarity data.

All methods, AP, RNG, and RGTM, have the drawback that they rely on the full dissimilarity matrix which is quadratic with respect to the number of data. Hence the techniques have quadratic complexity and they become infeasible for large data sets. Diverse approximations to get around a squared complexity in similar settings are available in the literature: kernel approaches can be accelerated to linear techniques by means of the Nyström approximation [28] which approximates the full Gram matrix by a low rank approximation. By integrating this approximation into the learning algorithms, an overall linear complexity results. We will show that the Nyström approximation can be extended to dissimilarity data and an integration into RNG and RGTM is possible. A linear time approximation results provided a fixed approximation quality of the matrix. We will show that, depending on the nature of the dissimilarity matrix, reasonable results can be obtained this way.

The Nyström technique has two drawbacks: it requires a representative set of examples for the low rank approximation, such that it cannot be used for online settings where data display a clear trend. Further, in order to arrive at linear techniques, the approximation has to be integrated into the learning algorithm. Hence this method does not constitute an option for clustering techniques where the entries of the dissimilarity matrix are used in a distributed way such as AP.

Patch processing has been proposed as an alternative approximation scheme. It offers a powerful linear time and limited memory approximation for streaming data sets [1]. In the article [15], it has been used to speed up RNG. The resulting technique, patch RNG (PRNG) is linear time. It requires a direct access to the dissimilarities. In this contribution, we transfer this technique to AP and RGTM, resulting in patch AP (PAP) and patch RGTM (PRGTM). We show that, depending on the given setting, good approximation quality can be achieved. Patch processing can even deal with streaming data which display a clear trend, unlike the Nyström approximation.

Now we first introduce the basic prototype based clustering techniques for general dissimilarity data. Then we introduce the basic idea behind the Nyström technique and patch processing and we show how these techniques can be applied to relational clustering. Finally, we compare the techniques using three benchmarks from bioinformatics: image data in the context of cytogenetics, mass spectra characterizing bacteria, and a part of the classical SwissProt database of protein sequences. In all cases, we compare the behavior of the techniques for data which are directly accessible form versus a presentation as streaming data. The tested algorithms, its approximations, and the corresponding acronyms are summarized in Table 1 for convenience.

Table 1
Different clustering algorithms and approximations introduced and tested in this contribution; algorithms which are suited for vectorial data only are set italic.

Algorithm	Acronym
Affinity propagation	AP
Patch affinity propagation	PAP
Neural gas	*NG*
Relational neural gas	RNG
Patch relational neural gas	PRNG
Nyström relational neural gas	RNG (Ny)
Generative topographic mapping	*GTM*
Relational generative topographic mapping	RGTM
Patch relational generative topographic mapping	PRGTM
Nyström relational generative topographic mapping	RGTM (Ny)

2. Prototype based clustering for dissimilarity data

We focus on prototype based clustering methods which represent clusters in terms of prototypical representatives. In the standard Euclidean setting, data $\vec{v}_i \in \mathbb{R}^n$ are represented by K prototypes $\vec{w}_j \in \mathbb{R}^n$. The receptive field R_j of a prototype $\vec{w}_j$ consists of all data points $\vec{v}_i$ which are closest to $\vec{w}_j$ as measured by the Euclidean distance $d(\vec{v}_i,\vec{w}_j) = \|\vec{v}_i - \vec{w}_j\|$, breaking ties deterministically. That means,

$$R_j := \{\vec{v}_i \in \mathbb{R}^n \,|\, d(\vec{v}_i,\vec{w}_j)^2 \le d(\vec{v}_i,\vec{w}_k)^2 \ \forall k \neq j\}.$$

In this setting, the goal of clustering can be formalized as minimizing the quantization error

$$E_{\mathrm{QE}} := \sum_{ij:\,\vec{v}_i \in R_j} d(\vec{v}_i,\vec{w}_j)^2, \tag{1}$$

to obtain prototypes which are as representative for their receptive field as possible. There exist different classical methods which achieve this goal: a direct optimization by means of a gradient descent as present in online vector quantization, or more advanced methods which take a neighorhood structure into account or which rely on a probabilistic interpretation of the model [19,10,4]. The latter techniques often yield much more stable results. Further, the possibly provide additional information such as the ability to visualize the prototypes such as in the generative topographic mapping or a neighborhood structure of the clusters such as in neural gas. We will consider three typical clustering techniques in this context: (I) The generative topographic mapping (GTM) which constitutes a generative statistical model. It models data by means of a constraint mixture of Gaussians induced by a mapping from a low-dimensional latent space. In latent space visualization is possible. (II) The neural gas (NG) which models data by means of representative prototypes which represent data in relation to the rank of its distance. This way a very robust algorithm is obtained which is widely independent from scaling issues. (III) Affinity propagation (AP) which reformulates the quantization error as a likelihood function. This can be decomposed as factor graph for which the max-sum algorithm can be used [10].