DSLib: An open source library for the dominant set clustering method

Sebastiano Vascon SEBASTIANO.VASCON@UNIVE.IT
University Ca' Foscari of Venice
Via Torino 155, 30172 Mestre (VE)

Vittorio Murino VITTORIO.MURINO@IIT.IT
Department of Pattern Analysis and Computer Vision
Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova , Italy

Marcello Pelillo PELILLO@UNIVE.IT
University Ca' Foscari of Venice
Via Torino 155, 30172 Mestre (VE)

Editor: ...

Abstract

DSLib is an open source implementation of the Dominant Set (DS) algorithm written entirely in Matlab. The DS methods is a graph-based clustering technique rooted in the evolutionary game theory that starts gaining lots of interests in the computer science community. It has been originally introduced in (Pavan and Pelillo, 2003, 2007) and, thanks to its duality with game theory and the notion of maximal clique, has been explored in many directions not only related to clustering problems. Applications in matching, segmentation, classification, medical imaging and network analysis are nowadays common in literature. In this package we propose not only the original implementation but a still growing collections, as more comprehensive as possible, of methods and modifications from different researchers based on the original core. Since an official implementation of the method has not been released the aim of this library is to fill this lack, making this repository of central role for this topic. For these reasons we decided to create a library easily integrable into a Matlab pipeline without dependences, simple to use and easily extendibles for upcoming works.
The latest source code, the documentation and some examples can be downloaded from https://xwasco.github.io/DominantSetLibrary/.

Keywords: Dominant set, clustering, graph, maximal clique, game theory

1. Introduction

Clustering is one of the still unsolved problem in the domain of pattern recognition. It is related to the task of grouping elements into meaningful subsets (e.g. points which are close in an Euclidean space) exploiting the underlying hidden structure of the data. Moreover clustering is an ill-posed problem since different partitioning of the same dataset could ends up into a meaningful subdivision of the data and thus the final results mostly depends on the function used to quantify the distances/similarities of points. Two properties char-

acterize in general every clustering algorithms: each subset (cluster) should have an *high internal coherence* and an *high external in-coherence* this means that items inside of the same cluster should have an high similarity while items of different clusters should have an high dissimilarity to achieves a good separation of the subsets. These two properties represent the foundation of many clustering algorithm and as well the *dominant set* (DS) method. DS has been formalized in the domain of graph theory with a strict relationship to the notion of weighted maximal clique, a set of mutually connected nodes in which the weight are mostly similar. A clique is a common representation for a cluster in the machine learning community. In the dominant set framework a generic dataset is represented as a weighted graph in which the vertices are the dataset's elements and the edges represents the similarity between pairs of nodes. The edges are usually weighted by a similarity function (the higher the value the closest the points). The algorithm iteratively find the dominant sets into this graph-based representation using a dynamical system that mimic a natural selection process. Cluster extraction are thus totally driven by the similarity of the elements in the graph. This shifts the traditional perspective of other clustering algorithm, in which the core is the minimization/maximization of a cost functional, to a stability condition of a dynamical system. Using this clustering technique is beneficial for many reasons and in particular when the following strong assumptions common in other clustering technique (like k-means, spectral clustering,etc.) cannot be satisfied. The DS can be employed as a clustering algorithm and is particularly suited when: i) the a-priori number of clusters to be found is totally unknown, ii) the pairwise similarity function return asymmetric affinities (is not a metric).

The rest of the paper is organized as follow: in Sec 1.1 a brief introduction to the dominant set method and evolutionary game theory are provided, in Sec2 the software package is explained with a practical samples and in Sec 3 and Sec 4 some case study and conclusions on the library are drawn.

1.1 Dominant Sets & Evolutionary Game-Theory

A dataset is represented in terms of an edge-weighted graph $G = (V, E, w)$ with no self-loop. The adjacency matrix of the graph G contains the relationship among the nodes in terms of similarity (higher value between pairs imply high similarity). Often, but is not mandatory, the matrix $A = (a_{ij})$ is constructed using a distance function through a Gaussian kernel $a_{i,j} = e^{-\frac{d(i,j)}{\sigma}}$ where $d(i,j)$ is a generic non-negative distance function between elements i and j while σ is a normalization term which set the decay of the negative exponential.
Finding a dominant set (cluster) is achieved optimizing this standard quadratic assignment problem:

$$\begin{aligned} &\max x^T A x \\ &\text{s.t. } x \in \Delta^n \end{aligned} \tag{1}$$

$$x_i(t+1) = x_i(t)\frac{(A\mathbf{x}(t))_i}{\mathbf{x}(t)^T A \mathbf{x}(t)} \tag{2}$$

where A is the similarity matrix and x is a n-dimensional vector lying in the n-dimensional symplex with $n = |V|$ the nodes in the graph. A solution can found iterating the Eq 2, a dynamical system known as *Replicator Dynamics* (RD), on the matrix A and on the

vector x until convergence. Other dynamical system are proposed in this package (Infection Immunization, Exponential Replicator Dynamics). RD has been designed in the context of Evolutionary Game Theory (Weibull, 1995) to model the changes in a population (vector x) based on the fitness of its individual (matrix A). RD act as a selection process over the population leading the fittest individual to survive (the more similar to each other) while the others to extinct. At convergence the support of $x, \delta(x) = \{i \text{ s.t. } x_i > \theta\}$ represents the set of the survived individual the, so-colled dominant set.

2. Description of the package

In this section we provide a short description of the library and how to integrate it in your pipeline. To get deeper into the subject check the papers and the inline documentation of the package.

2.1 Requisite

No particular requisite are needed, the library run on any Matlab environment (Windows, Linux, Mac) and it has no dependencies. It has been tested on Matlab 2014b and 2016b in Windows/Linux and MacOSX. Some components are written in C for optimization purpose but equivalent Matlab code are available without the needing of compiling them.

2.2 What's inside the package ?

In this version of the library we have implemented the following papers: Pavan and Pelillo (2003, 2007); Rota Bulo and Bomze (2011); Vascon et al. (2013) and a still growing set of papers will be added in the future.

2.3 Dynamical systems

The method uses a dynamical system to optimize the program (1) and thus for finding the dominant sets. The library comes with three different type of dynamics, *RepDyn, ExpRepDyn, InfImm(Rota Bulo and Bomze, 2011)*, and it can be easily extended to upcoming new optimization method for the program (1).

2.4 Pratical usage

In this section we show how to setup a basic running example to cluster a generated synthetic dataset.
Step 1: Generating data and build the affinity matrix A:

```
rng('default');                                % For reproducibility
cx = [1 1;5 5 ;8 8];                          %center of the clouds of points
npts=100; pts= repmat(cx,npts,1) + randn(npts*size(cx,1),2);
A=pdist(pts);                                  %pairwise Euclidean distances
s=3*var(A);     %an euristic to find sigma
A=exp(-A./s);  A=squareform(A); %from distance to similarity
A=A.*not(eye(size(A))); %the graph should not have self-loops
```

Step 2: Choosing the evolutionary dynamics and the DS parameters:

```
dynType=1;          %0=Replicator Dynamics, 1=InfectionImmunization ...
    2=Exponential replicator dynamics
precision=1e-6;  %the precision required from the dynamical system
maxIters=1000;  %number of maximum iteration of the dynamical system
x=ones(size(A,1))./size(A,1); %starting point of the dynamical system
theta=1e-5;     %threshold used to extract the support from x.
```

Step 3: Call the clustering method

```
[C]=dominantset(A,x,theta,precision,maxIters,dynType);
```

alternatively the clustering method can be called providing only the similarity matrix and using the default parameters.

```
[C]=dominantset(A);
```

Step 4: Show the cluster results

```
scatter(pts(:,1),pts(:,2),5,C); %show the points colored by cluster
```

The package comes with more complex examples and a deep inline documentation.

3. Case studies

The core described in this manuscript has been used in several studies. Here we highlight four recently published applications in radically different domains to prove the trasversality of the library and its effectiveness:

- **Protein Clustering**: In Pennacchietti et al. (2017) super-resolution images are processed in order to isolate group of molecules from the background noise. The molecules are given in the form of spatial localization (x,y) and the cluster to be found should respect certain constraints (number of points, densities, spreadness, etc.).
- **Conversational group detection**: In (Vascon et al., 2014, 2016) groups of people that are interacting are detected in images and video. Each person is characterized by a position and an head orientation. On top of these two features a third one is extracted based on sociological and biological constraints. This feature, called frustum, is extracted for each person and encoded into an histogram. The scene is thus represented as a graph in which the vertices are the persons and the likelihood of pair of person being in a group is computed using information-theoretic measures on the frustum. The dynamical system of the dominant set are used to extract the cliques from the graph.
- **Brain's fiber clustering**: In (Dodero et al., 2013) neuronal fibers of the brain are clustered for finding bundles connecting different anatomical areas. Bundles represent

higher level abstraction of the physical connection providing a way to perform tractography analysis without manual intervention. Each fiber is a vertex in a graph and the pairwise similarity matrix is generated based on the spatial and morphological similarity between pair of fibers. DS is applied over this representation extracting the bundles. In (Dodero et al., 2015) the above method has been extend including a matching phase across different subject bundles being able to recover common brain structures among subjects.

- **k-NN prototype selection**: In (Vascon et al., 2013) a k-NN classifier is boosted by finding representative prototypes using the dominant set. The training dataset is clustered and then labeled via majority vote. The clusters centroid are used to classify unseen items. This methods improved the performance of a standard k-NN classifier both quantitatively and in the timing since it uses only a small subset of sample during the classification.

4. Conclusion

Dominant set in the last decade has prooved its powerfulness leading to a variety of derived methods and applications. Here we present a library implementing the original paper and succeed and methods. The library is written in pure Matlab, is cross-platform, open source and does not have any dependency. Full documentation, tutorials and download can be found at https://xwasco.github.io/DominantSetLibrary/

Acknowledgments

We would like to acknowledge support for this project to the Pattern Analysis and Computer Vision departement at the Italian Institute of Technology, the University Ca' Foscari of Venice, the European Centre for Living Technology, Dr. Rota Bulo from Fondazione Bruno Kessler of Trento and Dr. Luca Dodero and prof. Diego Sona for testing the library and give us their valuable support.

References

L. Dodero, S. Vascon, L. Giancardo, A Gozzi, D. Sona, and V. Murino. Automatic white matter fiber clustering using dominant sets. In *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*, pages 216–219, June 2013.

Luca Dodero, Sebastiano Vascon, Vittorio Murino, Angelo Bifone, Alessandro Gozzi, and Diego Sona. Automated multi-subject fiber clustering of mouse brain using dominant sets. *Frontiers in Neuroinformatics*, 8:87, 2015. ISSN 1662-5196. doi: 10.3389/fninf.2014.00087. URL `http://journal.frontiersin.org/article/10.3389/fninf.2014.00087`.

M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–145–I–152 vol.1, June 2003.

Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):167–172, January 2007. ISSN 0162-8828.

Francesca Pennacchietti, Sebastiano Vascon, Thierry Nieus, Christian Rosillo, Sabyasachi Das, Shiva Tyagarajan, Alberto Diaspro, Alessio del Bue, Enrica Maria Petrini, Andrea Barberis, and Francesca Cella Zanacchi. Nanoscale molecular reorganization of the inhibitory postsynaptic density is a determinant of gabaergic synaptic potentiation. *Journal of Neuroscience*, 2017. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0514-16.2016. URL `http://www.jneurosci.org/content/early/2017/01/10/JNEUROSCI.0514-16.2016`.

Samuel Rota Bulo and Immanuel M. Bomze. Infection and immunization: A new class of evolutionary game dynamics. *Games and Economic Behavior*, 71(1):193–211, January 2011.

Sebastiano Vascon, Marco Cristani, Marcello Pelillo, and Vittorio Murino. Using dominant sets for k-nn prototype selection. In Alfredo Petrosino, editor, *Image Analysis and Processing ICIAP 2013*, volume 8157 of *Lecture Notes in Computer Science*, pages 131–140. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-41183-0.

Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. A game-theoretic probabilistic approach for detecting conversational groups. In *Asian Conference in Computer Vision*, Lecture Notes in Computer Science, 2014.

Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, 2016. doi: 10.1016/j.cviu.2015.09.012. URL `http://dx.doi.org/10.1016/j.cviu.2015.09.012`.

Jörgen W Weibull. *Evolutionary game theory*. The MIT press, 1995.