Biochimica et Biophysica Acta 1844 (2014) 2277–2283

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

On dimension reduction of clustering results in structural bioinformatics

Gábor Iván, Vince Grolmusz *

PIT Bioinformatics Group, Eötvös University, Pázmány Péter stny. 1/C, H-1117 Budapest, Hungary
Uratim Ltd., H-1118 Budapest, Hungary

ARTICLE INFO

Article history:
Received 29 January 2014
Received in revised form 23 August 2014
Accepted 27 August 2014
Available online 6 September 2014

Keywords:
Clustering
Protein sequences
Phylogenomics
Phylogenetics
OPTICS
SCOP classification
SCOP tree
SwissProt
UniProt
Sequence alignment

ABSTRACT

OPTICS is a density-based clustering algorithm that performs well in a wide variety of applications. For a set of input objects, the algorithm creates a *reachability plot* that can either be used to produce cluster membership assignments, or interpreted itself as an expressive two-dimensional representation of the clustering structure of the input set, even if the input set is embedded in higher dimensions. The focus of this work is a visualization method that can be applied for comparing two, independent hierarchical clusterings by assigning colors to all entries of the input database. We give two applications related to macromolecular structural properties: the first is a sequence-based clustering of the SwissProt database that is evaluated using NCBI taxonomy identifiers, and the second application involves clustering locations of specific atoms in the serine protease enzyme family—and the clusters are evaluated using SCOP structural classifications.

1. Introduction

Clustering algorithms are well-developed and frequently used data mining techniques that assign similar objects to the same group, creating several "clusters". The method is a core technique in data mining, can be used for knowledge discovery without any *a priori* hypothesis, and also has numerous applications in bioinformatical data mining of biological sequence data [1–6], in biomedical image processing and analysis [7], in proteomics data analysis [8], in microarray data analysis [9,10], protein–protein interaction network analysis [11,12] and phylogenomic analysis [13].

One of the most frequently used clustering algorithms is the k-means [14]. One step of the algorithm needs the computation of the center of gravity of the points clustered. Therefore, it can be applied mostly for datasets containing points in Euclidean spaces. While some versions of k-means overcome this constraint [15], the requirement of specifying the number of clusters prior the algorithm is run, or the property that the method prefers finding spherical clusters, is unwanted in numerous applications.

Some newer clustering methods apply non-geometric, statistical or a hybrid approaches [16–18]. In particular, for clustering protein sequences, the CD-HIT method is frequently applied [5,6]. The algorithm is built on the idea of hashing short subsequences of the input; consequently, it is very fast.

Some other clustering methods use only similarity measures (instead of distances) between any pair of objects, so they are more widely applicable than subsequent-hashing or center-of-gravity computing methods. Density based approaches, like DBSCAN [19] or OPTICS [20], are able to find non-spherical clusters as well. The interpretation of the output (and even properly setting the input parameters) of a given clustering algorithm is usually a non-trivial task.

For the visualization of high-dimensional clusters, researchers use either Johnson–Lindenstrauss random projections [21], principal component analysis, or some dimension-reducing non-linear mappings, like the Sammon-projection [22,23], or Self Organizing Maps (SOM) [24]. Here we suggest using the OPTICS [20] clustering method since it easily and naturally maps any high-dimensional cluster to the plane.

In order to visualize two different clusterings simultaneously, one needs to use very clear visualization of the clusters appearing in the distinct clusterings. Very little work was done in this direction. In some very simple and non-hierarchical cases a matrix visualization of joint distributions may give applicable results. For simultaneous embedding of two clusterings in the plane, some results are published in Ref. [25].

* Corresponding author at: Department of Computer Science, Eötvös University, Pázmány Péter 1/c, H-1117 Budapest, Hungary. Tel.: +36 1 3812226; fax: +36 1 3812231.
E-mail addresses: hugeaux@pitgroup.org (G. Iván), grolmusz@pitgroup.org (V. Grolmusz).

http://dx.doi.org/10.1016/j.bbapap.2014.08.015
1570-9639/

Hierarchical structures, that is, tree-structures, hierarchical clusters, multi-layer classifications are available in numerous field of biology, e.g.:

- The EC classification of enzymes http://www.chem.qmul.ac.uk/iubmb/enzyme/;
- all the phylogenetic trees, prepared by different methods, genetic materials and optimizations, are hierarchical structures that differ from one another (e.g., the Tree of Life based on mitochondrial DNA differs from the Tree of Life of constructed from the genomic DNA);
- the NCBI taxonomy database;
- the SCOP database of more than 1 million domain classifications http://scop.mrc-lmb.cam.ac.uk/scop/; and
- the CATH protein domain classification database of 16 million domains http://www.cathdb.info/.

In this work we propose a visualization method that makes use of hierarchically represented *a priori* knowledge available about the input objects, and assigns colors to them based on this information. We then show how the proposed method can help identifying clusters with the OPTICS clustering algorithm [20,26,27].

Any novel application of a bioinformatics method needs to be validated by detailed comparisons of known techniques and *a priori* knowledge (c.f. Refs. [28,29],). Our presented method yields a framework for this comparison: two independent clusterings can be superimposed: one clustering by the concave regions of OPTICS reachability diagram, and the other, completely independent clustering by the coloring of the data items in the reachability diagram. We applied this technique first in Ref. [26], where we found that in protease enzyme families, the configuration of just four spatial points in these enormous protein structures definitively implies their exact enzymatic role. In the present work, we formulate the visualization method itself.

Several methods are published for the numerical assessment of two different clusterings on the same set of data points. The classic measure is the Rand index [30], which gives the ratio of the number of pairs in the same class in both clusterings plus the number of pairs in different classes in both clusterings, compared to the number of all pairs of data objects. Some more recent methods evaluate the *quality* of some new or modified clustering method, compared to the known algorithms for more specialized problems [31,32,5,6]. For evaluating two different clustering of gene expression data, a Minimum Description Length-based method is proposed in Ref. [33]. An information-theoretical approach is described in Ref. [34]: the method measures the amount of information gained and lost by switching from one clustering to another. If both the gain and loss are small, then the clusterings are close; otherwise they are far from one another. A more involved statistical method is described in Ref. [35] for comparing hierarchical clusterings numerically.

Our method does not compute or introduce another numeric distance or measure of the similarity of cluster-analyses; we assign *colors* to the data points, based on one of the clusterings, and these colors are used to mark the data points in the second clustering. We prefer to use the OPTICS clustering method [20] for the visualizing the second clustering since the reachability diagram of OPTICS is always given in two dimensions, therefore, the visualization of the higher-dimensional clusters does not need projections or difficult transformations.

The source code of the visualization algorithm with sample output is available at http://uratim.com/appendix.zip.

2. Overview of the OPTICS clustering algorithm

Our visualization method proposed in the next sections can be applied to the output of any clustering algorithm. However, the usefulness of the method is going to be presented using results of the specifically chosen OPTICS algorithm [20], as the simultaneous use of OPTICS and the hereby presented visualization technique brings some further advantages. First we give a brief description of the OPTICS clustering algorithm, and also the justification of using this particular algorithm as a candidate to test our proposed visualization method.

For data clustering we intended to use an algorithm that is capable of identifying outlier points (also referred to as "noise") and is not biased towards even sized or regular shaped clusters. Density-based clustering algorithms have these desirable properties. The density of objects can be defined with a radius-like ϵ parameter and an object-count lower limit (*minpts*): a neighborhood of some object o is considered dense if there exist at least *minpts* objects within a less-than-ϵ distance. As the clustering structure of many real-data sets cannot be characterized by one (global) density parameter, it seems advisable to eliminate one of the above two input parameters and use it on the output instead.

The OPTICS (*Ordering Points To Identify the Clustering Structure* [20],) algorithm achieves this by *ordering* the objects contained in the database, creating the so-called *reachability plot*. The reachability plot is generated by assigning a value called *reachability distance* to all the objects of the database, while processing the objects in a specific order: the algorithm always chooses the object reachable with the smallest possible ϵ distance while maintaining the lower limit defined by *minpts*, meaning roughly the "most dense direction". This ensures that the hierarchical clustering structure of the database is also preserved.

The measure of local density for each object encountered is depicted on the reachability plot that contains almost all the information about the clustering structure of the database, although it does not directly assign the objects to clusters. There exist several methods that assign cluster memberships to objects based on the OPTICS reachability plot; these may be of interest in a future study. However – with the proposed visualization method – it is possible to obtain quite usable results without even assigning any particular cluster memberships to the objects: when using the OPTICS clustering algorithm together with a specific similarity measure, we would usually like to know whether the "deep" regions of the reachability plot – these are "potential" clusters – correlate with some *a priori*-known information.

The reachability plot of some points scattered on a two-dimensional plane is depicted in Fig. 1. The applied similarity measure is simply Euclidean distance. It is important to notice that the OPTICS algorithm is capable of creating the reachability plot for objects represented in arbitrary dimensions; it is only the similarity measure that has to be changed accordingly.

As a side effect, OPTICS reduces the dimensionality of the input dataset; combining OPTICS with the visualization method proposed later can be thus also used to visually compare two hierarchical clusterings of (possibly) multi-dimensional datasets. The literature of dimension reduction and visualization of high-dimensional data sets is quite rich (e.g. Ref. [36]), which is also true for visualizing hierarchical clusterings (e.g. Ref. [37]). Our method combines dimension reduction with visualization, making it possible to compare clustering results to an *a-priori* given hierarchical classification without assigning objects to specific clusters.

3. Coloring nodes of the *a priori*-known hierarchical data structure

In the visualization phase we are going to assign colors to each entry occurring on the x-axis of the OPTICS reachability plot, based on the *a-priori* given hierarchical classification of these objects. The main idea is that we would like to use similar colors on entries that belong to "similar" classes in the *a priori*-known hierarchical data structure. As this hierarchical structure can be conveniently represented by a (non-binary) tree (a *dendrogram*), our aim is to assign colors to tree nodes so that nodes having a short path between them (i.e. their common ancestor is close to them) are assigned similar colors. We would also like to achieve that the depth of a given node in this tree is somehow reflected in its color.

We will use the HSB (Hue, Saturation, Brightness) representation of colors. HSB coordinates can easily be converted to RGB (Red, Green, Blue) color coordinates. As an example, it is easy to see that points of