Temporal Cluster Analysis for radar satellite data

Gabriella Milone
University of Naples - Federico II
gamilone@unina.it

Abstract: *Clustering is a popular technique of data analysis and data mining. Since clustering problems are complex in nature, the larger is the size of the problem, the harder is to find the optimal solution and the longer it takes to reach reasonable results. Clustering techniques are conventionally divided in hierarchical and partitioning. In this paper I present a review of the clustering algorithms for large temporal databases and an application to radar satellite data in which I study different types of ground deformation trend by SAR images of the European Space Agency. The studied region is the area between the cities of Benevento and Avellino.*

Keywords: Temporal Clustering, Similarity, TDM, CLARA, Interferometry, PS-InSAR

1 Introduction

Gaining a relational understanding of information is important to biology, human cognition, artificial intelligence and many other data-intensive fields of research. In most cases finding a relationship is not obvious by inspection, given the high dimensionality of the data set. Hence computation is useful in order to divide the information into groups called clusters. To effectively divide the information it is necessary first to define a criterion for creating groups and, second, to find an optimal grouping based on that criterion.

In this paper I present a critical review of temporal clustering for large temporal databases supported by an application on real data from radar satellite measurements.

Clustering is a process of grouping data into a number of clusters, each of which contains the data that are similar to each other according to a specified similarity (distance) measure. A process of clustering is a sequential procedure which takes data (patterns) as a raw material and produces clusters as a result. Figure1 shows the steps of the clustering procedure.

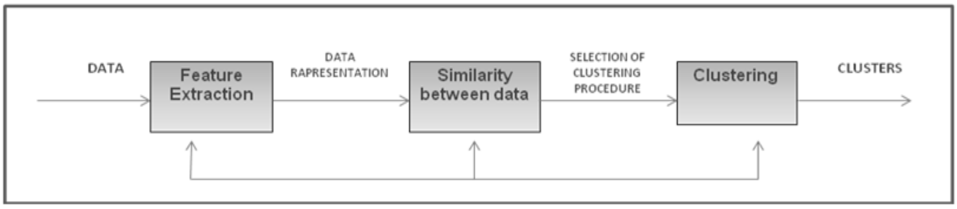


Figure 1: Clustering process

The most widely used similarity (distance) metric in clustering procedures is the euclidean distance, which is a particular case of the Minkowski metric with r=2:

$$d(x,y)=\left(\sum_{i=1}^{m}|x_i-y_i|^r\right)^{1/r} \rightarrow d(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \quad (1)$$

2 Temporal Clustering

Temporal Clustering is a particular case of clustering aimed at grouping a collection of time series (or sequences) based on their similarity. It is well known that the data clustering is inherently a more difficult task than the supervised classification, in which the classes are already identified so that a system can be adequately trained. But this intrinsic difficulty worsens if sequential data are considered, since the structure of the underlying process is often difficult to infer and the sequences are typically of different length.

Clustering of sequences or time series by similarity is a concept which appears in many disciplines (Figure 2).

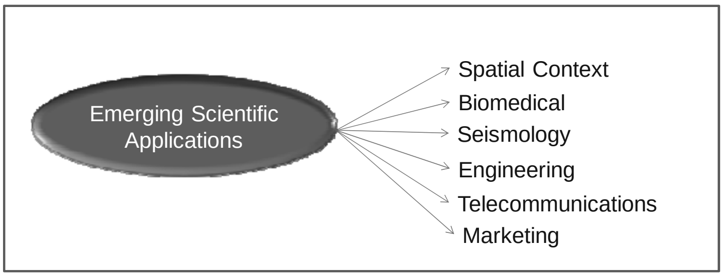


Figure 2: Scientific application in Temporal Clustering

The *quality* of a good temporal clustering method is measured by:

- high *intra-class* similarity
- low *inter-class* similarity.

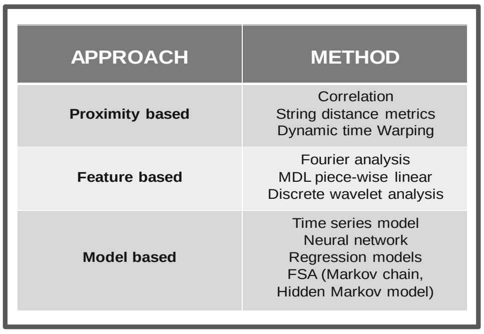
Furthermore, the *quality* of a temporal clustering result depends on both the similarity measure used by the method and its implementation.

A good clustering method is also measured by its ability to discover some or all of the *hidden* patterns.

If the number of clusters is given, then temporal clustering techniques can be divided into three groups:

- *Proximity-based*, where the main effort of the clustering process is devising the similarity or distance measures between sequences. With such measures, any standard distance-based method (as agglomerative clustering) can be applied
- *Feature-based* where the process extracts a set of features from each individual data sequence that captures temporal information. The problem of sequence clustering is thus reduced to an easier clustering (vector of features)
- *Model-based* where an analytical model is assumed for each cluster and the aim of the clustering procedure is to find a set of such models that best fits the data.

Temporal data clustering methods are summarized in Figure 3:



APPROACH	METHOD
Proximity based	Correlation String distance metrics Dynamic time Warping
Feature based	Fourier analysis MDL piece-wise linear Discrete wavelet analysis
Model based	Time series model Neural network Regression models FSA (Markov chain, Hidden Markov model)

Figure 3: Approach and methods of Temporal Clustering

3 Clustering Algorithms in temporal data

Clustering techniques are conventionally divided in hierarchical and partitioning. Hierarchical clustering is further subdivided into agglomerative and divisive. The basics of hierarchical clustering include the Lance-Williams formula, the idea of conceptual clustering and algorithms such as the now classic COBWEB or the newer CURE and CHAMELEON. I survey them in the Hierarchical Clustering section.

While hierarchical algorithms build clusters gradually (as crystals are grown), partitioning algorithms learn clusters directly. In doing so, they either try to discover clusters by iteratively relocating points between subsets, or they try to identify clusters as areas densely populated with data.

They are further categorized into:

- *probabilistic clustering* (EM framework, AUTOCLASS)
- *k-medoids methods* (PAM, CLARA, CLARANS, and its extension)
- *k-means methods* (different schemes, initialization, optimization, harmonic means, extensions).

Partitioning algorithms of the second type are surveyed in the Density-Based Partitioning section. They try to discover dense connected components of data, which are flexible in terms of their shape. Density-based connectivity is used in the algorithms DBSCAN, OPTICS, DBCLASD, while the algorithm DENCLUE exploits space density functions. These algorithms are less sensitive to outliers and can discover clusters of irregular shapes. They usually work with low-dimensional data of numerical attributes, known as spatial data. Spatial objects could include not only points, but also extended objects (algorithm GDBSCAN). Several algorithms work with data indirectly by constructing summaries over the subsets of the attribute space. They perform space segmentation and then aggregate appropriate segments. I discuss them in the Grid-Based Methods section. They frequently use hierarchical agglomeration as an intermediate step. The Algorithms BANG, STING, WaveCluster, and an idea of fractal dimension are discussed in this section. Grid-based methods are fast and they handle outliers well. Grid-based methodologies are also used as an intermediate step in many other algorithms (CLIQUE, MAFIA).

4 A temporal clustering application on radar satellite data

The ground deformations occurring in the central sector of the Campania Region has been investigated by applying the Permanent Scatterers Synthetic Aperture Radar Interferometry (PS-InSAR) on ERS1 and ERS2 satellites radar images dated between June 1992 and December 2000. The technique involves interferometric phase comparison of several radar images of the same scene (a portion of the earth surface, wide 100x100 km) taken at different times along the same orbit by the satellite radar sensors. Each satellite orbits at an elevation of 780 km and takes on the same image every 35 days. The output data have very high spatial resolution, allowing assessments of the motion of individual buildings and other man-made structures and of natural surfaces.

The processing technique, developed by the T.R.E. s.r.l. (a *POLIMI spin-off company*) is based on the identification of radar benchmarks, named *Permanent Scatterers* (PS), which are stable natural reflectors (rock outcrops, buildings and urban structures), characterized by stable individual radar-bright and radar-phase over long temporal series of interferometric SAR images.

The studied scene encompasses the area between the cities of Benevento and Avellino in the Campania region. The database delivered by the PS-InSAR analysis is formed by 18.452 PS characterized by a coherence value higher than 0.80 with a time series of 72 observations.

The time series of measured displacement allows us to identify regions in which there is a significant time dependent component to the deformation field and enables the deformation evolution of PS over time. In order to estimate ground deformation velocity trends, a statistical approach has been used: the

classification allows the identification of the different geological causes that produce a trend of deformation in a group of spatially coherent ground points.
The first step of Temporal Clustering consists in the application of different clustering algorithms (hierarchical and not) on the database.
The results of this application are the groups obtained by the CLARA (Clustering LARge Applications) algorithm that generates a composition of four best groups.
The choice of the algorithm is coherent from the standpoint of the interpretation of the results concerning the land deformation analysis. The main evidence is the presence of four main geological groups of different compositions (Figure 4).

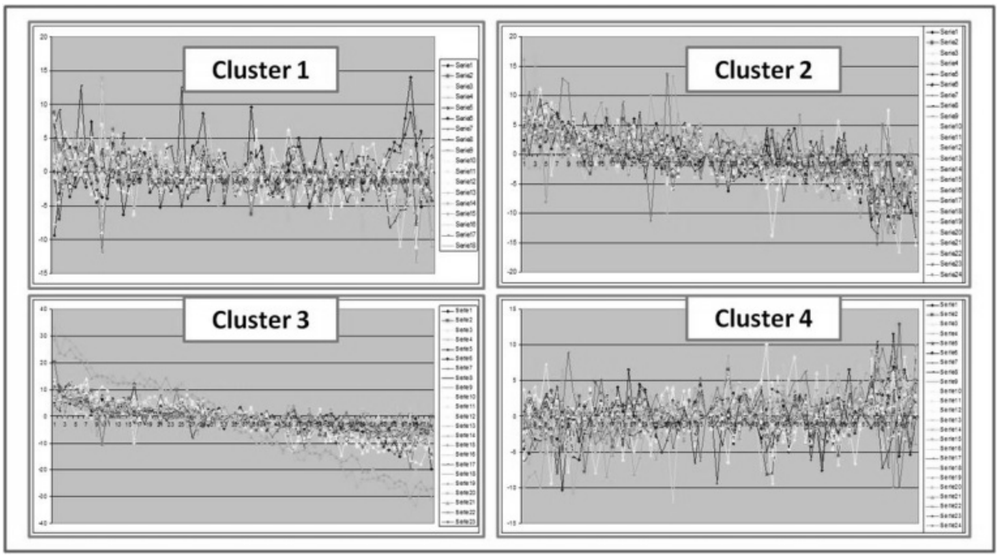


Figure 3: Rapresentation of four groups

Bibliography

Colesanti C., Ferretti A., Prati C., Rocca F. (2003), Monitoring landslides and tectonic motion with the Permanent Scatterers technique, Engineering Geology, 68: 3-14.

Farina P., Colombo D., Fumagalli A., Marks F., Moretti S. (2006), Permanent Scatterers for landslide investigations: outcomes from the ESA-SLAM project, Engineering Geology, 88: 200-217.

Ferretti A., Prati C., Rocca F. (2001), Permanent Scatters in SAR Interferometry, IEEE Transactions on Geoscience and Remote Sensing, 39:.8-20.

Kaufman L., Rousseeuw P.J. (1990), Finding Groups in Data: an Introduction to Cluster Analysis. Wiley and Sons.

Milone G. (2007), Temporal Data Mining: tecniche e algoritmi di clustering, Phd Thesis in Statistics, University of Naples Federico II.

Scepi, G. (2007), Clustering Algorithms for Large Temporal Data Sets. In: Proceedings of Cladag 2007, Springer, in press. Acknowledgements Many thanks to Antonio Risi (Representative PODiS of Regione Campania) and Carlo Terranova (Coordinator PODiS Campania Unit).