Journal of Computational and Applied Mathematics 344 (2018) 521–531

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

A novel scale-invariant, dynamic method for hierarchical clustering of data affected by measurement uncertainty

Federica Vignati *, Damiano Fustinoni, Alfonso Niro

Politecnico of Milano, Energy Department, via Lambruschini, 4 –20156 Milan, Italy

ARTICLE INFO

Article history:
Received 31 May 2017
Received in revised form 30 May 2018

MSC:
62H30
68T99

Keywords:
Hierarchical clustering
Non-uniqueness
Proximity measure
Computational cost
Uncertainty

ABSTRACT

An enhanced technique for hierarchical agglomerative clustering is presented. Classical clusterings suffer from non-uniqueness, resulting from the adopted scaling of data and from the arbitrary choice of the function to measure the proximity between elements. Moreover, most classical methods cannot account for the effect of measurement uncertainty on initial data, when present.

To overcome these limitations, the definition of a weighted, asymmetric function is introduced to quantify the proximity between any two elements. The data weighting depends dynamically on the degree of advancement of the clustering procedure. The novel proximity measure is derived from a geometric approach to the clustering, and it allows to both disengage the result from the data scaling, and to indicate the robustness of a clustering against the measurement uncertainty of initial data.

The method applies to both flat and hierarchical clustering, maintaining the computational cost of the classical methods.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

With the advancements in computer science, the study of clustering has gained increasing interest in the last decades, due to its several technical applications, ranging from machine learning [1], to data mining [2], including a number of tools for scientific research [3], logistics [4], and everyday life [5].

Indeed, in the study of many engineering and physical phenomena, data appear to be conditioned by several parameters. When analytical models are not available, the correlation between multivariate data and the corresponding parameter configurations can be usually obtained only by means of numerical simulations or experiments. Unfortunately, the nature of this correlation may remain unknown. Moreover, it is sometimes observed that experiments performed with different operational conditions may produce similar results. The first step to understand the data underlying structure, therefore, may consist in determining the families of configurations which lead to comparable results.

Since the first works, e.g., [6], statistical clustering has been used to provide a classification of a set of multivariate data, based on some similarities among the members of the same set (termed *cluster*). Flat clustering methods, in particular, identify a given number of sets and assign each configuration to one of them. Algorithms for flat clustering are usually very fast. On the contrary, the quality of the results depends in general on a-priori choices, e.g., the number of clusters: the research of the optimal number of clusters requires additional computations. Moreover, flat clustering methods cannot distinguish between "close" and "closer" configurations. These limitations are overcome by hierarchical clustering: the clusters are recursively identified and partitioned into sub-cluster, each characterized by an increasing degree of similarity among the

* Corresponding author.
E-mail addresses: federica.vignati@polimi.it (F. Vignati), damiano.fustinoni@polimi.it (D. Fustinoni), alfonso.niro@polimi.it (A. Niro).

https://doi.org/10.1016/j.cam.2018.05.062
0377-0427/© 2018 Elsevier B.V. All rights reserved.

belonging elements. Despite the slightly larger computational cost, therefore, hierarchical clustering methods retrieve a larger number of informations. Classical hierarchical clustering methods can be either agglomerative (or *bottom-up*) or divisive (or *top-down*): the first ones start out with as many clusters as the number of data – i.e., all the clusters are singleton sets – and, at each iteration, they gather them together until a single cluster is built, containing all data. On the contrary, the second ones start out with a whole set containing all data, and recursively partition it into smaller subsets until the selected level of detail. For complete introductions and descriptions of statistic clustering, the reader is addressed to [7,8].

Despite the better insight in the data structure provided, also hierarchical clustering methods suffer from some limitations. The resulting clustering, indeed, is non-unique, as it depends on the method used to quantify the similarity between its elements [9,10], on the treatment adopted to deal with special cases [11] and, mostly, on the data scaling [12–14]. To overcome the latter limitation, several solutions have been proposed so far. A first attempt to produce scale-invariant clusters [15] was based on the so-called Mahalanobis distance [16]. The latter is derived from the Euclidean distance, but it is made scale-invariant by introducing a normalization with respect to the data covariance. However, as pointed out by later works, the use of the variance to normalize data may result in meaningless clusters [17]. One of the most effective technique to find robust, scale-invariant clustering is based on a geometric approach, the so-called "minimum volume ellipsoids method", which allows to both disengage the results from the data scaling and to control the significance of the results [18–20]. Unfortunately, the computational efficiency of these geometric methods is usually lower than non-geometric ones.

To fill this gap, in this paper we introduce a novel, enhanced technique for hierarchical agglomerative clustering. The method is based on a geometric approach and, at the same time, on the definition of a weighted asymmetric distance function. The advantages of the adopted function are threefold: on the one hand, it allows to disengage the resulting clustering from the data scaling. On the other hand, thanks to the definition of the distance function, it requires the same computational effort as non-geometric methods. Eventually, the proposed technique intrinsically manages the measurement uncertainty, if any, even if the statistical distribution of the error is unknown. Therefore, this method can be successfully applied to both fully deterministic data and to experimental results affected by errors. When experimental results are analyzed, indeed, the choice of a suitable, robust clustering procedure is particularly relevant. On the one hand, the numerical values of the data, indeed, depend on arbitrary choices, e.g the units of measurement, the use of dimensional or dimensionless variables, the normalization of the results with respect to a reference value. An effective clustering method, therefore, must be scale-invariant. On the second hand, experimental data are usually affected by the measurement uncertainty: for this reason, a robust procedure should be recommended. Previous works on the clustering of non-deterministic data often require to know in advance the probability density function of the data [21–23], which may be unknown.

This work stems from the results of a long-term experimental campaign carried out at the ThermALab of Energy Department of Politecnico di Milano on the enhancement of heat transfer in forced convection of air flows through rectangular channels by means of square ribs in large variety of geometrical configurations. The Nusselt number and the friction factor – indicators of thermal and hydraulic performances, respectively – were experimentally measured. The ribs enhance heat transfer by periodically deflecting streamlines, interrupting boundary layer growth and destabilizing the flow. All these effects bring about an early transition to turbulent regime or promote turbulence. Unfortunately, to force a flow through a ribbed channel at a given Reynolds number, a larger pumping power is required, resulting in lower hydraulic performances with respect to a smooth channel in the same conditions. For each flow Reynolds number, a large number of rib geometries and configurations were tested, since the program was aimed at investigating the optimal rib configuration, i.e., producing the best compromise between the heat transfer enhancement and the induced hydraulic losses. The results showed a large dispersion of the data, and the apparent lack of an underlying criterion. For this reason, a cluster analysis was first performed on experimental data, in order to determine a possible structure, by means of classical clustering methods. Both plain data – i.e., on the Nusselt number and on the friction factor resulting from the experiments – and the same data normalized with respect to the Nusselt number and on the friction factor of the reference configuration – the smooth channel – were analyzed, in order to highlight both the absolute performances of each configurations and the difference with respect to the reference case. Unfortunately, the results of the clustering analysis of non-normalized and of normalized data were completely different. This fact represented an unexpected additional difficulty, since it is commonly enough to use non-dimensional number – i.e., Reynolds and Nusselt numbers and the friction factor – in order to provide unique correlations in most thermo-fluid-dynamical problems. On the contrary, the obtained clustering was not unique, highlighting a sensitivity problem. As observed also in other scientific researches dealing with different metrics [24], it can become a hard task to understand a priori whether non-normalized or normalized data should be used for the clustering. Moreover, classical clusterings proved to be not helpful in the perspective of determining the channel performances for two additional reasons. On the one hand, the effect of the measurement uncertainty was not known. On the other hand, classical methods forced to attribute the same relevance to both the performance indicators, whereas it is known that, depending on the technical applications, either the thermal or the hydraulic aspect must be privileged. To overcome these limitations, the novel method has been devised.

The paper is structured as follows: Section 2 reports the formulation of the method. In particular, the novel function adopted to measure the proximity between elements is defined, and its physical and geometrical interpretation is provided. An example of clustering obtained by means of the proposed method is provided in Section 3, including a comparison with the results of a classical clustering procedure. Conclusions and final remarks are reported in Section 4.

Download English Version:

https://daneshyari.com/en/article/8901777

Download Persian Version:

https://daneshyari.com/article/8901777

Daneshyari.com