DATA CLUSTERING PROCEDURES: A GENERAL REVIEW

A. Manuela Gonçalves[1*], António Gaspar-Cunha[2]

[1] Department of Mathematics, Centre of Mathematics, University of Minho; mneves@math.uminho.pt
[2] IPC-Institute for Polymers and Composites, University of Minho, Guimarães, Portugal; e-agc@dep.uminho.pt
[*] Correspondence: mneves@math.uminho.pt, Department of Mathematics, *Campus* de Azurém, University of Minho, 4800-058 Guimarães, Portugal

Key Words: *data science, multivariate data, clustering*

Abstract

In the age of data science, the clustering of various types of objects (e.g., documents, genes, customers) has become a key activity and many high-quality computer implementations are provided for this purpose by many general software packages. Clustering consists of grouping a set of objects in such a way that objects which are similar to one another according to some metric belong to the same group, named a cluster. It is one of the most valuable and used tasks of exploratory data mining and can be applied to a wide variety of fields. Research on the problem of clustering tends to be fragmented across pattern recognition, database, data mining, and machine learning communities. This work discusses the common techniques that are used in cluster analysis. These methodologies will be applied to data analysis in the framework of polymer processing.

Data clustering procedures

Human activities nowadays produce massive amounts of data and there has been a dramatic growth in the development of statistical methodology in the analysis of high-dimensional data.

Multivariate statistical analysis is concerned with analyzing and understanding data in high dimensions. The complexities of most phenomena require an investigator to collect observations in n individuals (number of subjects/examinees/individual or entities) on many different variables (p variables, that is, number of sets of measurements on a given individual). Multivariate analysis refers to all statistical techniques that simultaneously analyse multiple measurements on individuals or objects under investigation [1]. The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following:

1. Data reduction or structural simplification. The phenomenon being studied is represented as simply as possible without sacrificing valuable information. It is hoped that this will make interpretation easier;
2. Sorting and grouping. Groups of "similar" objects or variables are created, based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required;
3. Investigation of the dependence among variables. The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?
4. Modeling/Prediction. Relationships between variables must be determined for

predicting the values of one or more variables based on observations of the other variables;

5. Hypothesis construction and testing (statistical inference). Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

This study focuses on cluster analysis and describes a range of algorithms for investigating structure in data to find groups of objects or groups of variables that are more similar.

Clustering is one of the important data mining methods for discovering knowledge in multidimensional data [2]. The goal of clustering is to identify patterns or groups of similar objects within a data set of interest. In the literature, it is referred to as "pattern recognition" or "unsupervised machine learning" - "unsupervised" because we are not guided by a priori ideas of which variables or samples belong in which clusters. "Learning" because the machine algorithm "learns" how to cluster [3,4].

Therefore, to discover knowledge from a large amount of data, it is necessary to apply machine learning techniques, which are classified into two categories unsupervised methods and supervised methods:

- **Unsupervised methods**

These methods include mainly clustering and principal components analysis methods. The goal of clustering is to identify patterns or groups of similar objects within a data set of interest. Principal component methods consist of summarizing and visualizing the most important information contained in a multivariate data set. These methods are unsupervised because we are not guided by a priori ideas of which variables or samples belong in which clusters or groups. The machine algorithm "learns" how to cluster or summarize the data;

- **Supervised methods**

These methods consist of building mathematical models for predicting the outcome of future observations. Predictive models can be classified into two main groups: regression analysis for predicting a dependent variable. For example, you might want to predict life expectancy based on socio-economic indicators (independent variables); classification for predicting the class (group) of individuals. For example, you might want to predict the probability of being diabetes-positive based on the glucose concentration in the plasma of patients. These methods are supervised because we build the model based on known outcome values. That is, the machine learns from known observation outcomes to predict the outcome of futures cases.

There are different types of clustering methods, including partitioning clustering (no hierarchical, subdivides the data into a set of k groups), and hierarchical clustering (identify groups in the data without subdividing it). The classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations. The result of this computation is known as a dissimilarity or distance matrix. Dissimilarity/similarity represents the degree of correspondence among objects across all of the characteristics used in the analysis. It is a set of rules that serve as criteria for grouping or separating items: correlation measures (based-distance), and distance measures.

The choice of distance measures is very important, as it has a strong influence on the clustering results. For most common software, the default distance measure is the Euclidean distance. But it is also important to apply other distance measures: Squared Euclidean distance, Mean Euclidean distance, Weighted Euclidean distance, Minkowsky metrics, etc.

Depending on the type of the data and the researcher's questions, other dissimilarity measures might be preferred. If we want to identify clusters of observations with the same overall profiles regardless of their magnitudes, then we should apply the correlation-based distance as a dissimilarity measure.

Clustering variables that have scales widely differing numbers of scale points exhibit differences in standard deviations should be standardized. The most common standardization conversion is Z score (which means equal to 0 and a standard deviation of 1). So, before cluster analysis, it is recommended to scale the data, to make the variables comparable. This is particularly recommended when variables are measured on different scales; otherwise, the dissimilarity measures obtained will be severely affected.

The no hierarchical algorithms (partitioning clustering) that subdivide the data sets into a set of k groups require the number of clusters to be known or specified upfront, hard to tell what is the best value of the number of clusters to use. The process consists of transferring information between groups is to optimize certain conditions, i.e., until no object changes cluster. The more partitioning clustering are k-means, k-medoids, partitioning around medoids (PAM), and fuzzy analysis.

Hierarchical clustering is an alternative approach to partitioning clustering for identifying groups in the data set. It does not require pre-specify the number of clusters to be generated. The hierarchical algorithms can be divided into agglomerative and divisive: the hierarchical agglomerative algorithms start with the n initial objects, and each step, merge the closest pair of clusters until only one cluster is left; the hierarchical divisive algorithms start with one, all-inclusive cluster, and at each step, split a cluster until each cluster contains an object. The most usual methods of hierarchical agglomerative clustering are single linkage, complete linkage, average linkage, centroid method, median linkage, and Ward’s method.

The result of hierarchical clustering is a tree-based representation of the objects, which is also known as a dendrogram. Starting with each object as a separate cluster, the dendrogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster. Observations can be subdivided into groups by cutting the dendrogram at the desired similarity level. After constructing the dendrogram we define the cophenetic coefficient. The cophenetic coefficient is the correlation between the original dissimilarities and the cophenetic dissimilarities (cophenetic matrix associated with the dendrogram). For a high-quality solution, the magnitude of this value should be close to 1. For internal validation, the measures used are connectivity, silhouette width, and Dunn index. Finally, in defining adjacent clusters to the data in the study, it is important to know which or which differ from the variables that provided the grouping for cluster interpretation.

Applying clustering to polymer extrusion data

In this work, the above-mentioned clustering technics were applied to data generated during the multiobjective optimization of a single screw extruder with the aim of defining the optimal crew geometry considering simultaneously a Conventional Screw (CS) and a Maillefer Barrier Screw (MBS). Different case studies were analysed considering bi-objective optimization: Output (Q) versus Length for melting (L); Output (Q) versus Power consumption (Power); Output (Q) versus melt Temperature (T) and Output (Q) versus degree of mixing (WATS). Figure 1 shows the initial population and the non-dominated solutions of the final population for the case of Q versus WATS using Evolutionary Algorithms. As can be seen, the EA converges to both types of screws. These results can be explained by the knowledge about the process and are related to the capacities of each type of screw of accomplishing the objectives defined for the single

screw extrusion process, i.e., the simultaneous accomplishment of the refereed objectives.

The aim of this work is to apply the clustering techniques to the data generated during the optimization process, i.e., to different generations, to get some information about the evolution of the solutions during the optimization.

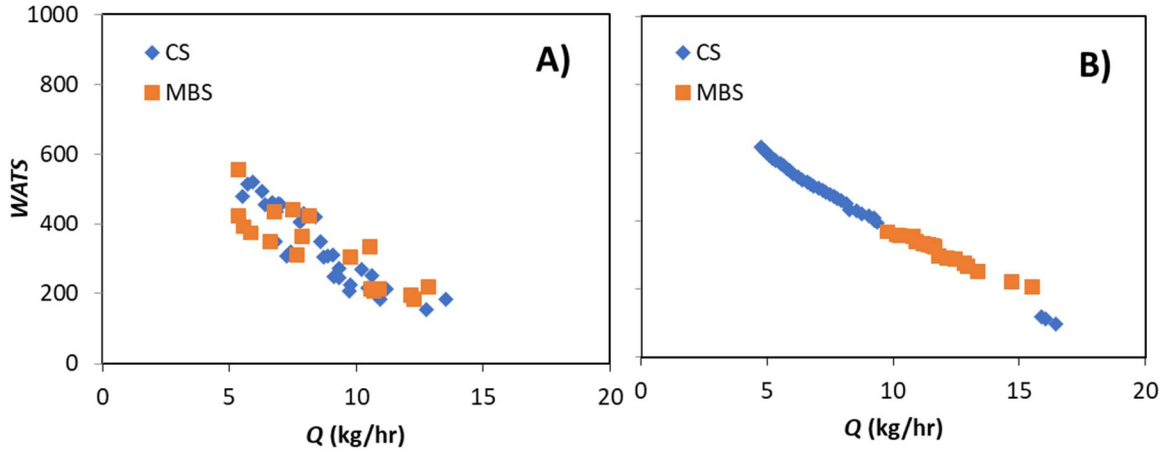

Figure 1- Optimization of CS and MBS using EAs: A) initial population and B) final non-dominated population.

Conclusions

The clustering methods applied proved to be effective (with high accuracy) for grouping the data accordingly to the importance of the decision variables and their effect on the objectives, i.e., in the performance measures.

REFERENCES

[1] Johnson, R.A., Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*. 6th edition, Prentice-Hall, Inc., New York.

[2] Aggarwal, C.C., Chandan, K.R. (2014). *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC.

[3] Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Edition1. Published by STHDA.

[4] Branco, A.J. (2004). *Uma Introdução à Análise de Clusters*. Sociedade Portuguesa de Estatística, Évora.

Acknowledgements

A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 734205 – H2020-MSCA-RISE-2016.