Input Selection of Merged Data in Manufacturing Process *

Mohamed Laib, Riad Aggoune, Eric Roseren

ITIS department, Luxembourg Institute of Science and Technology, Luxembourg
{mohamed.laib, riad.aggoune, eric.roseren}@list.lu

Keywords : *Metaheuristics, Sampling design, Design optimization, Industry 4.0, Data fusion.*

1 Introduction

Techniques for design optimization are often applied in various engineering domains to automatically enhance the performance of a component, or system or improve the quality of final products. The optimization of a complicated engineering system will certainly result in a problem with a wide variety of design parameters, which will need the use of complex computer simulations. Only a limited amount of data points are typically made available since the computational costs involved with running such simulations are very expensive [3]. Therefore, in most cases, two sources of data (collected from computer simulations and real experiments) are used in data-driven approaches, to predict product quality and/or construct a complete dataset [1].

From a data-driven point of view, merging two sources of data (data fusion) can be very challenging when some of the features do not have the same probability distribution or even the presence of clearly formed clusters based on the different sources. In this case, improving the available input space is required by selecting the best set of data points. In this work, we explore the possibility of employing evolutionary algorithms to choose the best set of data points that meet a particular criterion. The established approach is applied to a manufacturing case study, in which two data sources need to be merged in order to improve the prediction quality. The sampling is applied to data from both sources (simulations and real experiments). A well-dispersed data point will result from such an action, reducing the chance of producing clustered data while merging both datasets. The current strategy uses a global direct optimiser or explorer, such as Genetic Algorithm (GA), which may swiftly move away from places with poor fitness values while simultaneously studying a variety of attractive regions. [4].

2 Problem definition

The data used in this work comes from a manufacturing process. Two data sources are recorded from computer simulations and real experiments to assess the quality of the final product and understand in-depth the relationship with the input features. The main problem we are facing when merging these two datasets is the presence of two clusters (see Fig. 1). Since using both data sources is crucial, we need to perform sampling by considering some sampling criteria to ensure the homogeneity of data points. The simulated experiments compose a dataset of 300 individuals, while the dataset containing real experiments has 5000 data points.

*This research was funded by Luxembourg National Research Fund (Fond National de Recherche FNR), grant number BRIDGES18/IS/13307925.

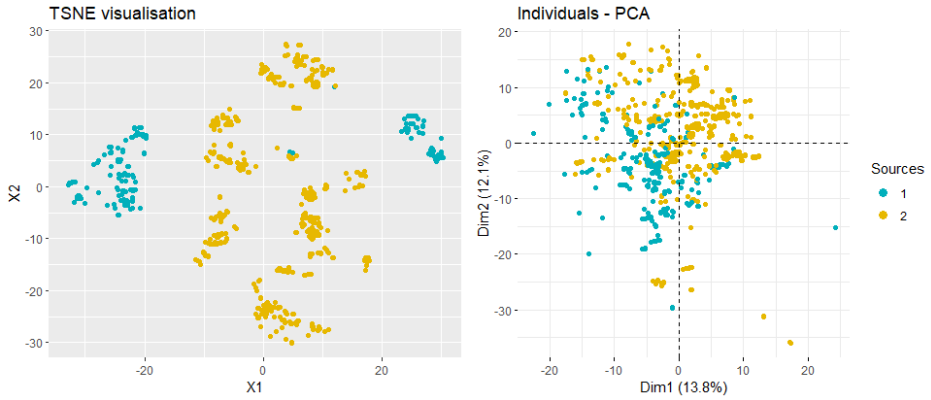

FIG. 1: Visualisation on new axes constructed using TSNE (left panel) and PCA (right panel).

3 Proposed Approach

The GAs are numerical techniques based on natural selection for resolving both constrained and unconstrained optimisation tasks. These algorithms repeatedly modify a population of individual solutions. The GA chooses individuals at random from the present population to be parents and uses them to produce offsprings for the following generation at each iteration or generation. The population adjusts to the best solution over successive generations.

Usually, the collection of one of the sources may be too costly, in this case, the most expensive source is used as a starting set of points and completed with the other one. The fitness measure implemented in the proposed approach is the coverage measure [2], which is originally used for sampling techniques. Using such a measure helps in selecting well-dispersed input space. The code implementing this approach will be available online during the workshop.

4 Conclusions et perspectives

The GA was used as global optimization to search for the best set of data-point between two data sources. In the present approach, the used fitness measure was the coverage measure. Other criteria from the design of experiments can also be used. Moreover, one can apply other criteria from other topics, such as clustering (e.g. silhouette measure). The proposed approach gave a new set of data less clustered than before, which can improve the prediction accuracy of the product quality.

During the workshop, other population-based methods will be compared. Moreover, comparing different criteria to use as a fitness measure will be presented to show which criterion is more suitable for the current manufacturing process.

References

[1] Marina Cocchi. Chapter 1 - introduction: Ways and means to deal with data from multiple sources. In Marina Cocchi, editor, *Data Fusion Methodology and Applications*, volume 31 of *Data Handling in Science and Technology*, pages 1–26. Elsevier. ISSN: 0922-3487.

[2] Mohamed Laib and Mikhail Kanevski. A new algorithm for redundancy minimisation in geo-environmental data. *Computers Geosciences*, 133:104328, 2019.

[3] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag.

[4] Michele Stramacchia. *Novel Ensemble of Surrogates-Based Infill Criterion for Engineering Design Optimisation*. PhD thesis, University of Southampton, November 2017.