A Genetic Algorithm for Feature Selection Applied to Data From Multiples Sources: Application to Manufacturing Data

Riad Aggoune, Mohamed Laib

Luxembourg Institute of Science and Technology, ITIS, Luxembourg
riad.aggoune@list.lu
mohamed.laib@list.lu

Abstract : *Reducing the redundancy in high dimensional data and finding the most relevant features is an important task in any data-driven approach. Especially, when the data consists of several datasets recorded from multiples sources. In fact, with such configuration, the redundancy can be within one source or even between different sources. This work explores a case study from manufacturing production process, in which, each step of production is considered as a source of data and contains many parameters (features). To reduce the dimensionality, an unsupervised feature selection has been applied using a genetic algorithm as search strategy.*

Keywords *: Dimension reduction, Genetic algorithms, Machine learning, Industry 4.0, Unsupervised learning.*

1 Introduction

As data volume is increasing rapidly, extracting knowledge is becoming more challenging. Especially when the data comes from multiples sources. For instance, meteorological data can be collected from satellites, ground-based sensors, and numerical models for forecasting. In genomics, in which Tang and Allen combined multiple genomic sources to gain insights into the genetic basis of Alzheimer's disease. They showed that the joint patterns extracted via integrated principal component analysis (iPCA) are highly predictive of a patient's cognition and Alzheimer's diagnosis [1].

This work aims at exploring manufacturing data collected from multiple sources in order to build predictive models for products quality. The case study describes the whole production procedure and each set contains a high number of features. Before building predictive models, the important step is to check if the input space is complete or redundant, in order to avoid difficulties in explaining the results [2]. Therefore, dimension reduction is needed for such kind of task. One of the methods that can be used is an unsupervised feature selection based on space-filling design [3]. However, the search strategy used for this method is a sequential forward search, which cannot be adapted for data collected from multiple source, because the resulted features strongly depend on the first selected one. In this work, a genetic algorithm is used as search strategy for this unsupervised feature selection method. Genetic algorithms were already used in feature selection for an accurate identification of informative features in multi-class and high-dimensional datasets [4]. The efficiency of using space-filling based feature selection implemented with a genetic algorithm is assessed first on simulated data and then on the real manufacturing case study. The obtained results outperform the previous configuration based on a sequential forward search.

2 Method description

The unsupervised feature selection used in this work consists of a filter algorithm based on space-filling design. More precisely, it is based on the coverage measure. In fact, the purpose is to reduce the redundancy among features and datasets collected in several steps. In other words, the goal is to find a set of features that fills the space in which the data is embedded. The coverage measure is defined as follows:

Let $X = \{x^1, \ldots, x^n\} \subset [0,1]^d$ be a sequence of n points of dimension d. The coverage measure is defined as follows:

$$\lambda = \frac{1}{\bar{\vartheta}} \left(\frac{1}{n} \sum_{i=1}^{n} (\vartheta_i - \bar{\vartheta})^2 \right)^{\frac{1}{2}} \tag{1}$$

where: $\vartheta_i = \min_{(k \neq i)} \left(dist\left(x^i, x^k\right)\right)$ is the minimal distance between x^i and the other points of the sequence. And: $\bar{\vartheta} = \frac{1}{n} \sum\limits_{i=1}^{n} \vartheta_i$ is the mean of ϑ_i; where *dist* is the Euclidean distance.

Genetic algorithm has 5 main steps: Generating populations, evaluating each chromosome in the population, selecting chromosome with best score, crossover chromosomes to reproduce new ones and mutate them with a low random probability [4]. In this work, the coverage measure is used as score (see Eq. 1)

3 Results and conclusions

The proposed methodology is applied on simulated data first to see its efficiency. The simulated data contains five datasets, each of them contains two independent features following a uniform distribution. Then, an additional four redundant features (constructed based on the uniform ones) are added to each dataset. The sequential forward search was not able to select the uniform features (non-redundant), because the first selected feature at the beginning. Regarding the results on the case study presented in this work, Table 1 shows the results of both search strategy, where we can clearly see that Genetic algorithm is not affected by the fact that data comes from multiple sources contrary to sequential forward search.

	Sel. Feat. (All feat.)	ROC AUC	S1	S2	S3	S4	S5
GA	65(293)	**0.78**	2(20)	6(33)	26 (115)	21 (97)	10 (28)
SFS	38(293)	0.72	4(20)	8(33)	9(115)	8(97)	8(28)

Table. 1: Results obtained using a Genetic Algorithm and Sequential Forward Search. The number of selected features vs the number of all feature between parentheses. The classification is performed using random forest on only selected features from all sources.

Further analysis can be performed on the results. For instance, estimating the intrinsic dimension of selected features from one source, in order to give more insights on information provided by each source as well as joint information between sources, since the selection does not depend on the first selected feature.

References

[1] T. M. Tang, G. I. Allen, Integrated Principal Components Analysis *Journal of Machine Learning Research*, 22(198):1 – 71, 2021.

[2] J. A. Lee, M. Verleysen, *Nonlinear Dimensionality Reduction.* 2007, Springer, New-York.

[3] M. Laib, M. Kanevski, A new algorithm for redundancy minimisation in geo-environmental data *Computers & Geosciences*, (133):104328, 2019.

[4] M. Chiesa, G. Maioli, G.I. Colombo, et al. GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. *BMC Bioinformatics* 21(54) 2020.