Bio-Inspired Algorithms for Feature Selection:
A Brief State of the Art

B

Rachid Kaleche
https://orcid.org/0000-0002-9326-2944
LIO Laboratory, Université Oran1, Algeria

Zakaria Bendaoud
https://orcid.org/0000-0003-3091-5044
GeCoDe Laboratory, University of Saida Dr. Moulay Tahar, Algeria

Karim Bouamrane
https://orcid.org/0000-0002-6953-0339
LIO Laboratory, Université Oran1, Algeria

Khadidja Yachba
https://orcid.org/0000-0002-7754-2823
LIO Laboratory, Université Relizane, Algeria

INTRODUCTION

Data is an important resource of knowledge used in decision making process, and which imply states, enterprises, and various organisations for different domains like economic, health, environment, policy, security, … Nowadays, the volume of data increases in an exponentially manner, and usual machine learning techniques reached their limits. According to the site Statista, the worldwide volume was 26 zetabytes in 2017, the forecasts for 2022 and 2025 are 97 zetabytes and 181 respectively. Therefore, extracting knowledge from voluminous data is a challenge due to the huge amount of features and/or the great number of instances. Hence, the high dimentionality implies a high computational cost, a difficulty in terms of readability due to the large amount of features, and irrelevant and redundant features could decrease the learning model performance (Dash, M., & Liu, H. 1997; Xue et al. 2016).

In order to reduce the high number of features, two main methods exist, which are feature selection (FS) and feature extraction (FE) (Kohavi, & John; 1997). The principal of FS is to select the relevant features, while in the case of FE, new features are constructed from the existant ones (Xue et al. 2016). This paper highlights on FS. The FS problem is a combinatorial problem, it is known to be an NP-Hard problem (Amaldi & Kann, 1998; Narendra & Fukunaga, 1977). The FS problem was addressed by classical algorithms such as sequential forward selection (SFS) and sequential backward selection (SBS) (Dash & Liu, 1997; Kohavi, & John; 1997) which reached their limits in terms of peformance and runtime (Xue et al. 2016). By contrast, these last decades global methods inspired from nature, namely bio-inspired algorithms, have been used and provided amazing results.

DOI: 10.4018/978-1-7998-9220-5.ch113

Nature have been an important inspiration source of algorithms to tackle hard problems, especially bio-inspired algorithms. These last have been inspired from living beings behaviors, as ants, bees, plants, bacteria and others, which smartly solve their own daily problems. Therefore, these bio-inspired algorithms have been used to address several hard problems in different fields as transport, medecine, robotic, industry and so on. In summary, bio-inspired algorithms proved their effectiveness and robusteness to tackle hard problems in a general manner. Since FS is a combinatorial problem, bio-inspired algorithms are, by consequence, the suited algorithms to tackle this problem. In addition, the hudge number of bio-inspired algorithms, more than 300 algorithms and their variantes, offers a wide range of choices.

Given that the FS process importance and the robustness and efficiency of the bio-inspired algorithms, the motivations of this paper are:

- Provide an overview of FS as an important process of machine learning to face the high dimentional issue of big data
- Provide a summary description of bio-inspired algorithms
- Describe bio-inspired algorithms modeling elements to address the FS problem
- Describe of bio-inspired approaches tackling FS problems
- Enumerate the application domains of bio-inspired algorithms for FS problems
- Outline challenges relevant to bio-inspired algorithms for FS problem, and those relied the bio-inspired algorithms in order to improve existant approaches tackling the FS process.

This paper is organized as follow: the second section is an overview of challenges related to the high dimentonality of data and the limits of classical FS algorithms. In the third section, an introduction to the bio-inspired algortihms is given. Some modelization elements of bio-inspired algorithms to address FS problem, and a brief state of art of applied bio-inspired algorithms with samples of applied approches are given in the forth section. In the fivth section, some future research directions are suggested. Finaly, the last section concludes this paper.

BACKGROUND

Big Data and Feature Selection

The term Big Data was introduced by Roger Mougalas from O'Reilly Media in 2005. The data are collected from several sources as internet, digital devices, IoT, … The data are characterised by a set of features, which could be of a few number or around thousands even millions of features like those collected by microarray technology for instance.

To face the challenges of high dimentionality of data, there is an issue to reduce search space and in the same time preserving a high performance (accuracy) of learning algorithms. These challenges are two opposite objectives. In addition, the reduction of the number of features leads to a better readability of the induction factors. There are two categories of methods to achieve the reduction of search space and keeping a high induction performance, namely FS and FE (Kohavi & John, 1997; Qiu & Feng, 2016). FS is the process which aims to select a subset of relevant features that will be used in a learning (induction) algorithm. In other words, the FS process aims to remove irrelevant and redundant features. On the other hand, FE is the preprocess that is used to construct a new subset of features from the original