Swarm and Evolutionary Computation xxx (2018) 1–13

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

A constructive evolutionary approach for feature selection in unsupervised learning

Nádia Junqueira Martarelli, Marcelo Seido Nagano *

Industrial Engineering Department, São Carlos School of Engineering, University of São Paulo, 400 Trabalhador São-carlence avenue, 13566-590, São Carlos, São Paulo, Brazil

ARTICLE INFO

Keywords:
Genetic algorithm
Feature selection
Clustering problem
Constructive Genetic Algorithm

ABSTRACT

In this paper, a novel Constructive Genetic Algorithm (CGA) for the feature selection in clustering problem is addressed. This issue has become a challenge since the data sets dimension increased exponentially over the years. In order to evaluate the CGA performance, the Genetic Algorithm (GA) has also been executed to be compared to the first one. The modeling and execution of this evolutionary approach to this problem are unpublished in the literature. For the results emission, twelve data sets have been used, of which four were simulated and eight are real data sets. The results showed that both approaches overperformed the no feature selection data sets. However, the CGA presented a better performance than GA in eight of the twelve data sets regarding solution quality. Considering the execution time, the CGA obtained exceptional results, that is, it spent less time than the GA in most data sets.

1. Introduction

By 2003, the humanity had already generated five exabytes of data. After twelve years (2015), the same exabyte amount is reached every 2 years, and this number is increasing [1]. Nowadays, data are easily and quickly obtained, with variability and in large volumes. The data sets size, number of instances, and dimension, number of attributes, are also increasing. Many non-relevant attributes are present in them, disturbing a possible knowledge discovery.

For that, a process, called Knowledge Discovery in Database (KDD), was organized into three main steps that guide the computational process of KDD in a large volume of data, where in the fist step occurs the data dimensionality reduction, eliminating non-relevant attributes.

The evolution of data sets dimensionality is shown in Fig. 1. The data sets are from the three repositories, namely: (i) University of California Irvine machine learning repository (UCI); (ii) University of California Irvine machine learning repository for Knowledge Discovery in Databases (UCI KDD) and (iii) Library for Support Vector Machines (LIBSVM).

In 1988, the largest dimensionality data set (soybean) contained 35 attributes, whereas in 2010, the biggest one (KDD2010) contained 29.890.095 [2].

Despite a considerable amount of information, not all attributes have contributed to knowledge discovery. In many cases, instead of helping, some of them may distort the results. A good example is given by Liu and Motoda [3], Fig. 2.

The authors show a formation of three groups resultant from two relevant attributes (COL1 and COL2), Fig. 2 (a). They also bring, in Fig. 2 (b), two irrelevant attributes (COL 3 and COL4), forming no groups. However, if three attributes are considered, Fig. 2 (c) and 2 (d), there is no group as defined as the previous one (2 dimensions).

Selecting only relevant attributes is a combinatorial optimization problem, with 2^p as the space of solutions, where p is the number of attributes, characterizing an NP-hard problem. Regarding the feature selection in clustering problem, the space of solution increases, because the number of groups is unknown. For that, in each subset there are $[k_{max}, k_{min}]$ possible groups, making the space of solution equal to $2^p[(k_{max} - k_{min}) + 1] - 1$, subtracting the empty subset $\{\emptyset\}$. In this case, there is a Bi-Objective Optimization Problem, because while it is necessary to find the best subset of attributes it is also necessary to find the best number of groups.

A good way to solve this kind of issue is to apply heuristics methods, such as Genetic Algorithm (GA),to achieve a good solution, in an adequate time since it is impossible to find an optimal solution executing

* Corresponding author.
E-mail address: drnagano@usp.br (M.S. Nagano).

https://doi.org/10.1016/j.swevo.2018.03.002
Received 15 May 2017; Received in revised form 18 December 2017; Accepted 2 March 2018
Available online XXX

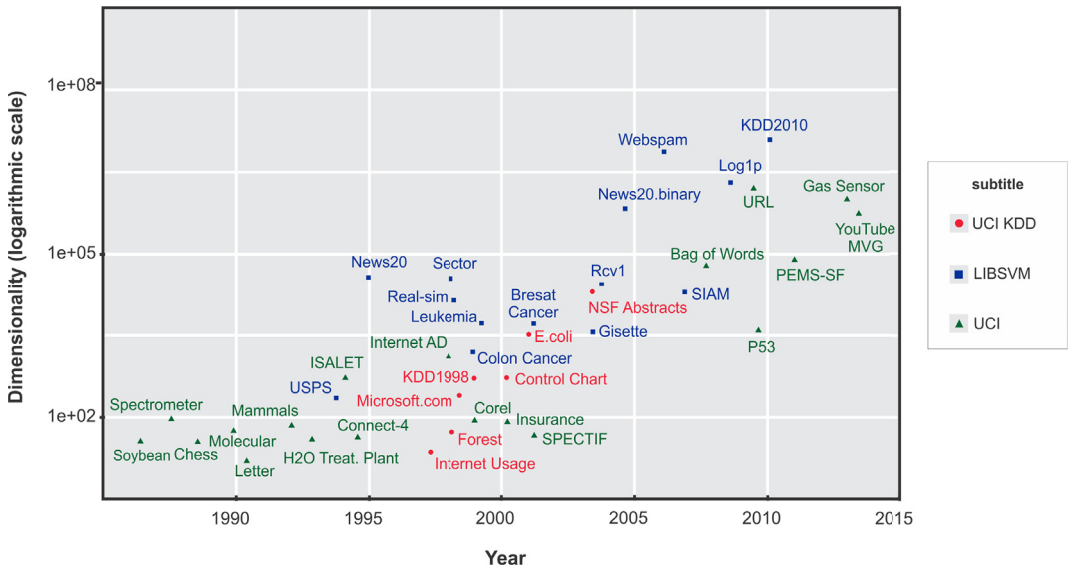

Fig. 1. Dimensionality evolution in data sets inserted in three repositories (i) UCI KDD, in red (ii) UCI, in blue and (iii), LIBSVM, in green, in the last two decades [2]. Adapted by author. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

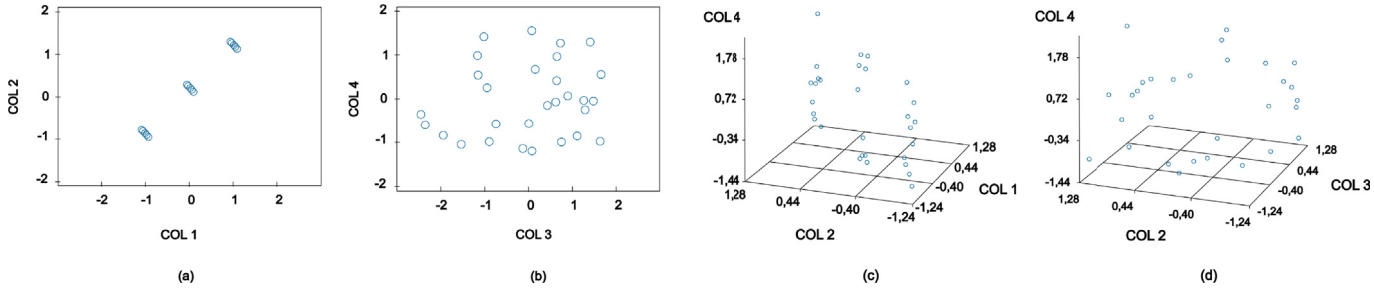

Fig. 2. Data plotted of different subset of attributes [4]. Adapted by author.

an exhaustive search due to high computational time.

The GA has been created by Holland [5] and has been inspired in natural evolutionary processes. Despite its incontestable success, the genetic algorithm requires improvements. For that, some branches were developed, one of them is the Constructive Genetic Algorithm (CGA). This alternative approach obtained relevant results in similar combinatorial optimization problems, such as simulated annealing and tabu search [6].

In this paper, the modeling and the application of CGA in clustering problem for feature selection are described in details. A comparison between CGA and GA is done in order to evaluate the CGA performance. Seven data sets are used, four of them have been created by Gaussian and uniform distribution and three others are real data set, namely banknote authentication, white wine quality, and air quality, which were extracted from the University of California Irvine machine learning repository [7–9]. This work is the first to apply the CGA for selecting attributes in clustering problem. No research works with this theme have been found in literature.

This paper is organized as follow. Section 2 presents briefly the clustering data theory. Section 3 the GA and the CGA theory are explained and the modeling details of these approaches are addressed. Section 4 reports the computational experiments. Section 5 brings the results and discussion. Finally, Section 6 presents the conclusion.

2. Clustering data

For Jain and Dubes [10], a clustering can be defined as an aggregation of points in a test space, wherein the distance between any two points in a group is less than the distance between any point in the group and any point outside the group. Many fields have been using clustering techniques to pattern recognition, customer segmentation, and trend analysis. Databases with numeric, categorical, and mixed attributes may be accepted by clustering algorithm depending on the metric used to measure the distance between the points. An example is the K-means algorithm with Euclidean distance, which takes only numeric data.

This algorithm divides the n records of the data set into k mutually exclusive groups, where $k <= n$. This clustering happens by randomly choosing k centroids in the domain of the set, which will be the initial representative of the group. The other records of the set are assigned to the groups by the distance they are from the centroids [11].

The K-means pseudocode are displayed in Algorithm 1.

Algorithm 1 Clustering algorithm K-means, [12]. Adapted by author.

```
Define the number of groups, k;
Assign each record to a group;
Evaluate the population;
while there are records changing groups do
    Calculate the central point for each group;
    Rearrange the records so that each one belongs to the nearest group;
end
```