Computers & Operations Research 37 (2010) 1381 – 1388

Contents lists available at ScienceDirect

Computers & Operations Research

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

Investigation of a new GRASP-based clustering algorithm applied to biological data

Mariá C.V. Nascimento*, Franklina M.B. Toledo, André C.P.L.F. de Carvalho

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Caixa Postal 668, São Carlos-SP, CEP 13560-970, Brazil

ARTICLE INFO

Available online 5 March 2009

Keywords:
Clustering
GRASP
Gene expression data
Bioinformatics

ABSTRACT

A large amount of biological data has been produced in the last years. Important knowledge can be extracted from these data by the use of data analysis techniques. Clustering plays an important role in data analysis, by organizing similar objects from a dataset into meaningful groups. Several clustering algorithms have been proposed in the literature. However, each algorithm has its bias, being more adequate for particular datasets. This paper presents a mathematical formulation to support the creation of consistent clusters for biological data. Moreover, it shows a clustering algorithm to solve this formulation that uses GRASP (Greedy Randomized Adaptive Search Procedure). We compared the proposed algorithm with three known other algorithms. The proposed algorithm presented the best clustering results confirmed statistically.

1. Introduction

In the last years there has been a considerable growth of the amount of biological data available in several domains. Two of these domains are proteomics, which studies properties found in proteins, and transcriptomes, which analyzes the transcripts, by measuring the level of mRNAs. One of the studies in proteomics is protein identification through the spectrometry mass from the peptide sequencing or from its mass fingerprinting. Proteins are complex macromolecules with 3D structures. The structure of a protein shows the traits of its functionality. However, protein structure prediction is a combinatorial problem [1]. Transcriptome can be studied by several methods, able to measure the mRNA levels, including gene expression data produced by microarray experiments. Analysis of gene expression data allows the discovery of meaningful groups of genes with related functionalities.

The use of clustering algorithms to discover new and useful information in biological data is getting increasing attention lately. Clustering algorithms are considered a powerful tool for the identification of groups and sub-groups in biological data. Clustering algorithms aim to group data consistently, in such a way that the most similar objects belong to the same group or cluster and dissimilar objects are assigned to different clusters. The use of these algorithms allows to detect similar objects in a dataset that could not be easily or efficiently grouped by humans. Cluster analysis has been applied to several domains, like natural language processing [2], galaxy formation [3] and image segmentation [4]. Surveys and reviews on clustering algorithms and their application to different domains can be found in [5,6]. In particular, a large number of experiments using clustering algorithms to group biological data have been published [7–9].

Clustering algorithms can be roughly divided into two main approaches: hierarchical and partitioning algorithms. The hierarchical approach produces a nested series of partitions. According to the type of the algorithm, it agglomerates and/or divides previous clusters to produce the final partition. The partitioning approach produces the partition that optimizes a given criterion or objective function. Despite the large number of algorithms proposed for each approach, each algorithm has its bias, being better suited for particular data distributions.

Recently, there has been a growing interest in the use of metaheuristics for data clustering. Several metaheuristics are adopted by partitioning clustering algorithms [8,10–13]. Metaheuristics are systematic algorithms able to produce and to find well valued solutions. Some of the main metaheuristics includes: Tabu Search, Simulated Annealing, Greedy Randomized Adaptive Search Procedure (GRASP) and Genetic Algorithms. To our knowledge, GRASP has never been applied to cluster biological datasets. This paper presents a novel GRASP for clustering based ona mathematical model. GRASP is a search algorithm that builds initial solutions through a semi-greedy process, and applying a local search around each solution previously built. It was proposed in [14] and has been successfully used in many combinatorial optimization problems [15–19]. Cano et al. [17] proposed a GRASP for clustering problem. The initial solution is built using the Kaufman greedy initialization

* Corresponding author. Tel.: +55 16 3373 9691; fax: +55 16 3373 9751.
E-mail addresses: mariah@icmc.usp.br (M.C.V. Nascimento), fran@icmc.usp.br (F.M.B. Toledo), andre@icmc.usp.br (A.C.P.L.F. de Carvalho).

0305-0548/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2009.02.014

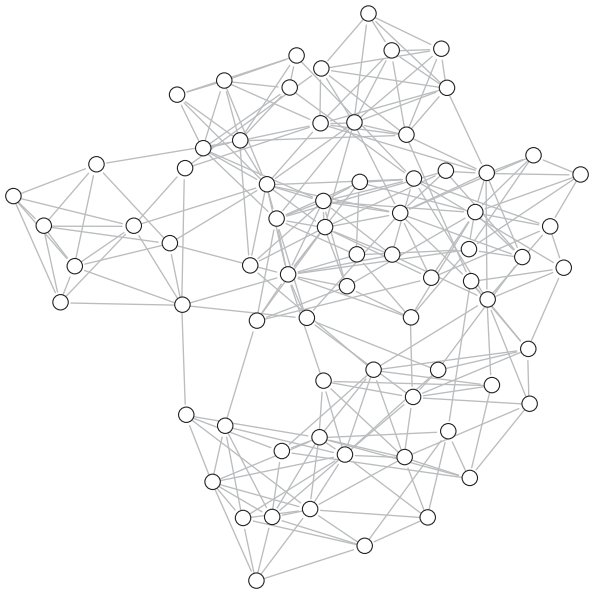

Fig. 1. Graph representation of a dataset.

[20]. Then, the local search and the K-means algorithm are performed.

In order to show how the proposed GRASP works, we use a graph representation of the datasets. The graph representation, as well as clustering algorithms based on graph theory, has been successfully applied in the past [21–23]. Fig. 1 illustrates a graph representation of a dataset. The nodes represent the objects whereas the edges indicate the relationship between objects.

To verify if the proposed metaheuristic performs well for clustering problems, we present a linearization of the studied mathematical model and use it to run some experiments in CPLEX [24]. According to the experimental results, CPLEX took a long time (some hours) to find optimal solutions for datasets with more than 40 objects and GRASP found the same solutions as CPLEX in less than 1 s. These results indicate that the proposed GRASP has a good potential, validating GRASP performance for the analyzed mathematical model. For all biological datasets used in the experiments, GRASP found the optimal solution in less than 1 min.

Furthermore, we accomplished some tests to assess the performance of the proposed algorithm using external knowledge. This experiment aims to evaluate the performance and the suitability of the model to cluster data. We carried out experiments and compared the results obtained by GRASP with the partitions produced by three known algorithms. The measure employed to evaluate all algorithms was the Corrected (adjusted) Rand index, named CRand [25]. This index evaluates the partitions with respect to the real classification provided for the datasets. Our purpose is to show that the GRASP, a metaheuristic based on the optimization mathematical model, is an efficient technique for clustering biological data. The tested datasets are related to diverse biological nature, such as protein fold classification, prediction of localization protein sites and cancer diagnosis.

Moreover, some of these datasets were clustered into different structures in order to increase the number of clustering problem investigated. We also showed the performance of the algorithm for a benchmark dataset, the Iris dataset [26]. The results show that GRASP achieved the best performance for the majority of the datasets.

2. Mathematical model

The mathematical programming has been successfully used to solve clustering problems [27–30]. Rao [28] presented many mathematical models in order to analyze different clustering approaches. As a special case, the author presented a model whose objective is to minimize the total within group distances. In this paper, we investigate this mathematical model for data clustering. The formulation, with a slightly modification from the original model, is given by

$$\min \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} \sum_{k=1}^{M} x_{ik} x_{jk}$$

subject to

$$\sum_{k=1}^{M} x_{ik} = 1, \quad i = 1, \ldots, N \tag{1}$$

$$\sum_{i=1}^{N} x_{ik} \geqslant 1, \quad k = 1, \ldots, M \tag{2}$$

$$x_{ik} \in \{0, 1\}, \quad i = 1, \ldots, N, \ \ k = 1, \ldots, M \tag{3}$$

where d_{ij} is the distance between objects i and j; N is the number of objects; M is the number of clusters; x_{ik} is a binary variable that assumes value 1, if the object i belongs to the cluster k and 0, otherwise.

The objective function aims to minimize the distance between the objects inside the same cluster. Constraints (1) assure that object i belongs to only one cluster. Constraints (2) guarantee that cluster k contains at least one object inside it. Finally, constraints (3) assure that the variables x_{ik} are binaries.

This model is characterized by its nonlinear objective function. It is necessary to use nonlinear integer programming to solve it, since there is no polynomial algorithm able to find the optimal solution. In the attempting to solve the problem using an efficient optimization software, we presented a linearization of the previous model. The linearized model is

$$\min \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} y_{ij}$$

subject to (1)–(3)

$$y_{ij} \geqslant x_{ik} + x_{jk} - 1, \quad i = 1, \ldots, N, \ \ j = i+1, \ldots, N, \ \ k = 1, \ldots, M \tag{4}$$

$$y_{ij} \geqslant 0, \quad i = 1, \ldots, N, \ \ j = i+1, \ldots, N \tag{5}$$

where y_{ij} is a real variable that assumes the value 1 if the objects i and j belong to the same cluster.

The differences between the former and the latter mathematical models are the objective function and constraints (4) and (5). Nevertheless, both objective functions have the same interpretation, since they minimize the distance between all objects that belong to the same cluster. Constraints (4) and (5) guarantee that y_{ij} assumes the value 1 if both values of x_{ik} and x_{jk} are 1. This linear model has $N^2/2$ more variables and $N(N-1)(M+1)/2$ more constraints than the nonlinear modeling. To solve it, we used CPLEX, an optimization solver that, roughly explaining, uses the branch and bound technique and many preprocessing operations to find optimal solutions of a mixed integer programming problem.

3. Proposed algorithm

In order to solve the proposed mathematical model for clustering, we investigate a GRASP-based algorithm. To our knowledge, there is no GRASP metaheuristic applied to this mathematical model. GRASP is a metaheuristic that consists basically of a constructive phase and a local search phase. In the constructive phase, the procedure builds an initial solution through a semi-greedy process. Each step of the