Expert Systems with Applications 41 (2014) 5277–5284

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Hybrid heuristics based on column generation with path-relinking for clustering problems

Rudinei Martins de Oliveira [a], Luiz Antonio Nogueira Lorena [a,*], Antonio Augusto Chaves [b], Geraldo Regis Mauri [c]

[a] National Institute for Space Research – INPE, São José dos Campos, SP 12227-010, Brazil
[b] Federal University of São Paulo – UNIFESP, São José dos Campos, SP 12231-280, Brazil
[c] Federal University of Espírito Santo – UFES, Alegre, ES 29500-000, Brazil

ARTICLE INFO

Keywords:
Hybrid heuristics
Column generation
p-Median problems
Clustering problems

ABSTRACT

This paper examines hybrid heuristics for solving clustering problems. The clustering problem can be defined as the process of separating a set of objects into groups such that members of a group are similar to each other. The methods are based on the application of a column generation technique for solving *p*-medians problems. Five heuristics are derived directly from the column generation algorithm: a solution made feasible from the master problem, the column generation solution, a heuristic with path-relinking considering the initial columns of the column generation procedure, a solution of the master problem with path-relinking and the column generation process with path-relinking. Solutions are tested with the external measure CRand and the computational results compared to recent methods in literature.

1. Introduction

The clustering problem is defined as the process of separating a set of objects into groups such that members of a group are similar to each other (Lorena & Furtado, 2001). The determination of similarity between individuals depends on the metric defined for the distance between objects. External measures are used to measure the similarity between sets, such as the Rand (Rand, 1971) and correct adjusted Rand (CRand) indexes (Hubert & Arabie, 1985).

The difficulty in solving the clustering problem comes from how to identify close objects and develop methods that fulfill the function of grouping them. This is not an easy task due to the great difficulty in finding good solutions, i.e., classify objects efficiently. The classification of data can be supervised or unsupervised (Abbasi & Younis, 2007). In supervised analysis the method is trained with known patterns of data to use it in new samples. In the unsupervised analysis, the algorithm seeks to find the data structures that allow the separation into groups without prior knowledge of the standards. Although methods of unsupervised classification get less accurate results than supervised methods, they are more suitable when no prior information about the groups is available.

The clustering problem examined in this paper is considered unsupervised, i.e., concerned with the grouping of related objects without information from class or label (Nascimento & De Carvalho, 2011). The clustering problem has been applied in a wide variety of research areas such as machine learning, artificial intelligence, pattern recognition, spatial data mining, image segmentation, genetics, microbiology, geology, remote sensing, among others (Xu & Wunsch, 2005; Jun, Park, & Jang, 2014). Authors have applied metaheuristics for resolution, such as greedy randomized adaptive search procedure (Nascimento, Toledo, & de Leon Ferreira de Carvalho, 2010) and genetic algorithms (Agusti'n-Blas et al., 2012).

This work proposes five hybrid heuristics to solve clustering problems. All of them are based on the application of a column generation technique for solving *p*-median problems (Senne, Lorena, & Pereira, 2007). The five approaches are: a solution made feasible from the master problem, the column generation solution, a heuristic with path-relinking considering the initial columns of the column generation procedure, a solution of the master problem with path-relinking and the column generation process with path-relinking. All clustering solutions are tested with the external measure CRand and the computational results compared to recent methods in literature.

The paper is organized as follows. Section 2 gives a brief literature review on the clustering problem. Section 3 presents an overview of column generation for *p*-median problems. Section 4

* Corresponding author.
E-mail addresses: rudmart@gmail.com (R.M. de Oliveira), lorena@lac.inpe.br (L.A. Nogueira Lorena), antonio.chaves@unifesp.br (A.A. Chaves), mauri@cca.ufes.br (G.R. Mauri).

http://dx.doi.org/10.1016/j.eswa.2014.03.008
0957-4174/

describes the hybrid heuristics for solving the clustering problem. Section 5 presents the data, the distances and correlations used to calculate the dissimilarity between samples, and the computational results. In Section 6 some conclusions are mentioned.

2. Abbreviated literature review

The clustering problem has been extensively studied. Rand (1971) proposes patterns that isolate aspects of performance of a method, such as, return, sensitivity and stability. These patterns depend on a similarity measure between two different clusters of the same set of data; the measure should essentially consider how each data point pair is assigned in each set.

Handl, Knowles, and Kell (2005) show the large amount of techniques available for validation of results obtained for the problem, with the main focus of the work toward the application of post-genomics data analysis. The authors use synthetic and real biological data to demonstrate the benefits and also some validation risks.

The work of Lorena and Furtado (2001) presents a Constructive Genetic Algorithm for solving the clustering problem. Mitra and Banka (2006) introduced a multi-purpose evolutionary biclustering with local search strategies. Hong, Kwong, Chang, and Ren (2008) propose an unsupervised selection algorithm called clustering ensembles guided feature selection algorithm, which combines clustering method with a learning algorithm. Chang, Zhang, and Zheng (2009) propose a separation algorithm based on genetic algorithms with gene rearrangement to *k*-means clustering, which seeks to remove the degeneracy for the purpose of a more efficient search. A crossover operator that exploits a measure of similarity between chromosomes in a population is also presented.

Recently, Nascimento et al. (2010) present a mathematical formulation and a greedy randomized adaptive search procedure to solve the clustering problem using biological data. The computational results were compared with the direct application of CPLEX (ILOG, 2009) solver, *k*-means, *k*-medians and Partitioning Around Medoids (PAM). The CRand index is used to compare the methods. Serpa (2011) applied an iterated local search heuristic combined with heuristic Clustering Search to the same data and have some improvements in CRand indexes.

The column generation approach for clustering problems has been little explored in the literature. Senne and Lorena (2001) and Senne et al. (2007) describe some relationships between the surrogate and Lagrangean relaxation techniques embedded in column generation for *0–1* linear programming problems. The combined Lagrangean/surrogate relaxation with column generation accelerates and stabilizes the primal and dual bounds obtained through the selection of reduced costs. In Filho and Lorena (2010), a heuristic based on column generation was applied to the cell formation problem of machines and parts. The paper presents a new model as a partitioning problem with cardinality constraint.

3. Column generation for *p*-median problems

The heuristics to solve clustering problems start from a set of data and without any information of standards build groups that have similar characteristics among its objects. The groups are obtained by a column generation technique proposed to solve *p*-median problems. *P*-median solutions minimize the sum of distances between nodes and their nearest facility (median) and it is expected that the characteristics between nodes that are allocated to the same facility are similar. Thus, a feasible solution to the clustering problem would result from *p*-median solutions, i.e., the separation of a set in *p* distinct subsets, with each group containing a facility and the nodes allocated it.

Given a distance matrix $[d_{ij}]_{n\times n}$, according to Senne et al. (2007), the *p*-median problem can be modeled as the following set partitioning problem with a cardinality constraint.

Minimize:

$$\sum_{k=1}^{m} c_k y_k \tag{1}$$

subject to:

$$\sum_{k=1}^{m} A_k y_k = 1 \tag{2}$$

$$\sum_{k=1}^{m} y_k = p \tag{3}$$

$$y_k \in \{0,1\} \tag{4}$$

Where the set $S = \{S_1, S_2, \ldots, S_m\}$ is formed by subsets of objects in $N = \{1, \ldots, n\}$; $M = \{1, 2, \ldots, m\}$ is the set of indices corresponding to columns of subset S; $A_k = [a_i]_{n\times 1}$, $k \in M$, with $a_i = 1$, if $i \in S_k$, and $a_i = 0$ otherwise; $c_k = Min_{i\in S_k}\left(\sum_{j\in S_k} d_{ij}\right)$ for $k \in M$ and y_k are the decision variables, with $y_k = 1$ if the subset S_k is chosen and $y_k = 0$ otherwise. For each set S_k, the choice of the median is performed by calculating the cost (c_k).

As the number of columns may be very large, the problem to be solved is a linear programming relaxation of (1)–(4) known as Master Problem (MP) and defined as follows:

Minimize:

$$\sum_{k=1}^{m} c_k y_k \tag{5}$$

Subject to:

$$\sum_{k=1}^{m} A_k y_k \geqslant 1 \tag{6}$$

$$\sum_{k=1}^{m} y_k = p \tag{7}$$

$$y_k \in [0,1] \tag{8}$$

Then, after defining a set of initial columns, the MP is solved, and their final dual costs (μ_i, for $i = 1, \ldots, n$) and ρ are used to generate new columns ($\alpha_j = [\alpha_{ij}]_{n\times 1}$), solving the following subproblem:

$$Min_{j\in N}\left[Min_{\alpha_{ij}\in\{0,1\}}\sum_{i=1}^{n}(d_{ij}-\mu_i)\alpha_{ij}\right] \tag{9}$$

The subproblem (9) is solved by considering each $j \in N$ as a median, and $\alpha_{ij} = 1$, if $(d_{ij} - \mu_i) \leqslant 0$ and $\alpha_{ij} = 0$ if $(d_{ij} - \mu_i) > 0$. For the new set S_j, define the vertices i such that $\alpha_{ij} = 1$ for the subproblem (9). Then, the column $\begin{bmatrix}\alpha_j \\ 1\end{bmatrix}$ is added to MP if the solution value of subproblem (9) is less than ρ. All columns found that satisfies inequality (10) for $j = 1, \ldots, n$, can be added to the set of columns, accelerating the column generation process.

$$\left[Min_{\alpha_{ij}\in\{0,1\}}\sum_{i=1}^{n}(d_{ij}-\mu_i)\alpha_{ij}\right] < \rho \tag{10}$$

The column generation algorithm is summarized in the flowchart of Fig. 1. After defining an initial set of columns, the software CPLEX (ILOG, 2009) is used to solve MP and obtain the dual values $\mu_i, i = 1, \ldots, n$ and ρ. They are employed to generate new columns ($\begin{bmatrix}a_j \\ 1\end{bmatrix}$) through the solution of the subproblem (9). All columns ($\begin{bmatrix}a_j \\ 1\end{bmatrix}$) which satisfy the inequality (10) (for $i = 1, \ldots, n$) are incoming