A Discrete Artificial Bees Colony Inspired Biclustering Algorithm

R. Rathipriya, Periyar University, India
K. Thangavel, Periyar University, India

ABSTRACT

Biclustering methods are the potential data mining technique that has been suggested to identify local patterns in the data. Biclustering algorithms are used for mining the web usage data which can determine a group of users which are correlated under a subset of pages of a web site. Recently, many blistering methods based on meta-heuristics have been proposed. Most use the Mean Squared Residue as merit function but interesting and relevant patterns such as shifting and scaling patterns may not be detected using this measure. However, it is important to discover this type of pattern since commonly the web users can present a similar behavior although their interest levels vary in different ranges or magnitudes. In this paper a new correlation based fitness function is designed to extract shifting and scaling browsing patterns. The proposed work uses a discrete version of Artificial Bee Colony optimization algorithm for biclustering of web usage data to produce optimal biclusters (i.e., highly correlated biclusters). It’s demonstrated on real dataset and its results show that proposed approach can find significant biclusters of high quality and has better convergence performance than Binary Particle Swarm Optimization (BPSO).

Keywords: Bee Colony Optimization (BCO) Algorithm, Biclustering, Correlation Based Fitness Function, Discrete BCO, Mean Squared Residue (MSR) Score, Web Usage Mining

1. INTRODUCTION

In the literature, clustering is the most commonly used data analysis technique. Standard clustering methods (such as K-Means, hierarchical clustering and self organizing maps) are partitions the set of objects into distinct groups called clusters based on their similarities existing across the features. Thereby, they may fail to uncover clusters that are similar only over some but not all features. In contrast, biclustering aims at finding subsets of users which are behaving similarly over a subset of pages. The usefulness of this concept in the context of microarray measurements has been demonstrated in different studies (Busygin et al., 2002; Cheng et al., 2000). In the context of microarray analysis, biclustering was firstly considered by Cheng and Church in 2000. Cheng and Church (CC) algorithm (Cheng et al., 2000) is a greedy iterative search method and consists in building a bicluster adding or removing rows or columns iteratively, thus, improving its quality which is measured with the Mean Squared Residue (MSR). The MSR is based on the sum of the squared residues which measure how adequate
each expression value is, in comparison with the rest of the values of the bicluster. In Getz et al. (2000), an iterative hierarchical clustering is separately applied to each dimension and biclusters are built by means of the combination of the obtained results for each dimension.

There is a group of biclustering algorithms based on meta-heuristics such as evolutionary approaches (Bleuler et al., 2004; Divina et al., 2006), multiobjective evolutionary approaches (Banka et al., 2006; Divina et al., 2006), Simulated Annealing (Bryan et al., 2005), Particle Swarm Optimization, greedy randomized adaptive search (Dharan et al., 2009), Estimation of Distribution Algorithms (Liu et al., 2006) or Memetics Algorithms (Gallo et al., 2009). All these algorithms used the MSR as a part of their fitness function. The MSR is effective in recognizing biclusters with shifting patterns but not some patterns with scaling trends, in spite of representing quality patterns. Aguilar-Ruiz et al. (2009) proved that the MSR is not a good measure in order to discover patterns in data when the variance of values is high, that is, when the users present scaling patterns. And a very few biclustering work has been done in the field of web usage mining. Xu et al. (2010), presented a co-clustering algorithm using bipartite spectral clustering to extract bicluster from web users and pages and the impact of using various clustering algorithms is also investigated in that paper. A novel web co-clustering algorithm named Co-Clustering in Semantic space (COS) was proposed by Zong et al. (2010), which simultaneously partitions web users and pages via a latent semantic analysis approach.

Among the various biclustering methods proposed in the literature, most approaches are based on greedy heuristics that refine a set of biclusters iteratively. These algorithms can be considered as local search methods which are fast but often yield suboptimal results. An improvement is possible by integrating the traditional biclustering methods (such as Two-Way clustering, Greedy Biclustering and etc.) into a global search strategy, e.g., an evolutionary algorithm (EA), optimization algorithms and etc. The contributions of this paper are

- Propose a general bio-inspired biclustering algorithm that can be coupled with a Two-Way clustering method
- Apply and investigate the proposed biclustering approach on web usage datasets

Bees Colony optimization and other black-box optimization methods have been used for standard clustering but to our best knowledge not in the context of biclustering.

The remainder of the paper is organized as follows. Section 2 briefly discusses about the methods and materials for the proposed biclustering algorithms. Section 3 introduces the Discrete Artificial Bee Colony (DABC) based biclustering algorithm and discusses its implementation aspects of web usage data used in this study. Experimental results are discussed in Section 4, and the last section is devoted to concluding remarks.

2. METHODS AND MATERIALS

2.1. Biclustering

Biclustering tries to identify local sub patterns in an arbitrary data matrix. In this context, biclusters are nothing but the local patterns of users’ browsing behavior, i.e., subsets of users that behave similarly over a subset of pages. Accordingly, a bicluster is defined as a pair \((U,P)\) where \(U \subseteq \{1,\ldots,m\}\) is a subset of users and \(p \subseteq \{1,\ldots,n\}\) a subset of pages; it can be easily shown that the space \(X\) of all biclusters given a matrix \(A\) is exponential in both \(m\) and \(n\), where \(X = \{1,\ldots,m\} \times \{1,\ldots,n\}\).

The optimization goal is to find one or more biclusters that are optimal with respect to their homogeneity (i.e., Average Correlation Value) and their size (Volume). These two criteria are usually competing, and the various biclustering approaches in the literature differ in the way how the conflicting goals are formulated and combined.
A Research Roadmap of Big Data Clustering Algorithms for Future Internet of Things
[www.igi-global.com/article/a-research-roadmap-of-big-data-clustering-algorithms-for-future-internet-of-things/223188?camid=4v1a](www.igi-global.com/article/a-research-roadmap-of-big-data-clustering-algorithms-for-future-internet-of-things/223188?camid=4v1a)

A Case Study on Data Quality, Privacy, and Evaluating the Outcome of Entity Resolution Processes