Knowledge-Based Systems 56 (2014) 108–122

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data ☆

Peter Peng [a], Omer Addam [a], Mohamad Elzohbi [a], Sibel T. Özyer [b], Ahmad Elhajj [c], Shang Gao [a], Yimin Liu [a], Tansel Özyer [d], Mehmet Kaya [e], Mick Ridley [c], Jon Rokne [a], Reda Alhajj [a,f,*]

[a] Department of Computer Science, University of Calgary, Calgary, Alberta, Canada
[b] Department of Computer Engineering, Cankaya University, Ankara, Turkey
[c] Department of Computing, University of Bradford, Bradford, UK
[d] Department of Computer Engineering, TOBB University, Ankara, Turkey
[e] Department of Computer Engineering, Firat University 23119, Elazig, Turkey
[f] Department of Computer Science, Global University, Beirut, Lebanon

ARTICLE INFO

Article history:
Received 16 April 2013
Received in revised form 22 September 2013
Accepted 1 November 2013
Available online 14 November 2013

Keywords:
Clustering
Genetic algorithm
Gene expression data
Multi-objective optimization
Cluster validity analysis

ABSTRACT

Clustering is an essential research problem which has received considerable attention in the research community for decades. It is a challenge because there is no unique solution that fits all problems and satisfies all applications. We target to get the most appropriate clustering solution for a given application domain. In other words, clustering algorithms in general need prior specification of the number of clusters, and this is hard even for domain experts to estimate especially in a dynamic environment where the data changes and/or become available incrementally. In this paper, we described and analyze the effectiveness of a robust clustering algorithm which integrates multi-objective genetic algorithm into a framework capable of producing alternative clustering solutions; it is called Multi-objective K-Means Genetic Algorithm (MOKGA). We investigate its application for clustering a variety of datasets, including microarray gene expression data. The reported results are promising. Though we concentrate on gene expression and mostly cancer data, the proposed approach is general enough and works equally to cluster other datasets as demonstrated by the two datasets Iris and Ruspini. After running MOKGA, a pareto-optimal front is obtained, and gives the optimal number of clusters as a solution set. The achieved clustering results are then analyzed and validated under several cluster validity techniques proposed in the literature. As a result, the optimal clusters are ranked for each validity index. We apply majority voting to decide on the most appropriate set of validity indexes applicable to every tested dataset. The proposed clustering approach is tested by conducting experiments using seven well cited benchmark data sets. The obtained results are compared with those reported in the literature to demonstrate the applicability and effectiveness of the proposed approach.

1. Introduction

A multi-objective genetic algorithm based clustering method is described in this paper. Its applicability and effectiveness are demonstrated by using some benchmark datasets, mainly related to gene expression data analysis which constitutes a vital research area of social and scientific impacts. Fortunately, clustering is one of the key methods that can be employed to the benefit of the computational biology and bioinformatics research communities. It allows researchers to identify molecules that demonstrate similar behavior or characteristics and hence could lead to utilizing in the analysis reduced set of molecules by considering representatives from each cluster instead of the whole original set of molecules.

In general, existing clustering techniques require prespecification of the number of clusters or some parameters that indirectly lead to the number of clusters; and these are is not an easy to predict *a prior* even for experts. Thus, the problem handled in this paper may be articulated as follows: Given a set of data instances (we mainly concentrate on gene expression data), it is required to develop an approach that produces different alternative solutions, and then rank the resulting solutions by conducting validity analysis. In fact, there are always some trade-offs between

☆ This paper is part of the project sponsored by Scientific and Technical Research Council of Turkey (Tübitak EEEAG 109E241). Tansel Ozyer would like to thank TUBITAK for their support.

* Corresponding author.
E-mail address: alhajj@cpsc.ucalgary.ca (R. Alhajj).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.11.003

the quality of a clustering result and the number of clusters. One solution is to view the two elements as two objectives that affect clustering results, i.e., this is naturally a multi-objective optimization problem. The solution of a multi-objective optimization problem is a set of alternatives which in one way can be seen as a Pareto-optimal set or non-dominated set [52].

In general, traditional algorithms for clustering microarray data do not produce the Pareto optimal set, and they do not lead to the optimal number of clusters in the database that they work on. For example, the hierarchical clustering method can get the heuristic overview of a whole dataset, but it cannot relocate objects that may have been 'incorrectly' grouped at an early stage. It can neither tell the optimal number of clusters nor give the non-dominated set. Partitional clustering like *K*-means needs the number of clusters as a predefined parameter, and it may lead to local optimal solutions because it concentrates on a local search from a random initial partitioning. SOM has the same disadvantage in that it requires the number of clusters as *a prior*. Clearly, a more advanced and comprehensive clustering algorithm is needed to get the global pareto-optimal solution set required to give users the best overview of the whole dataset according to the number of clusters and their quality. Further, it is required to get clustering results with the optimal number of clusters.

Clustering different samples based on gene expression is one of the key issues in problems like class discovery, normal and tumor tissue classification, and drug treatment evaluation [1,69]. Scherf et al. [58] applied microarray analysis on the gene expression database for the molecular pharmacology of cancer. It contains 728 genes, 60 cell lines, and 15 cell line groups. Golub et al. [17] applied SOM clustering algorithm on gene expression data containing 38 acute leukemia samples and 50 genes after filtered the whole dataset. SOM automatically grouped the 38 samples into two classes with acute myeloid leukemia (ALL) and acute lymphoblastic leukemia (AML). They further used SOM to group the samples into four classes. Subclasses of ALL, namely, B-lineage ALL and T-lineage ALL were distinguished [17]. It has been indicated that clustering samples can be used to identify fundamental subtypes of any cancer [58].

Clustering analysis can also be used to find direct gene-sample correlations. BiCluster [13] enables Gene/Condition correlation analysis that can lead to molecular classification of disease states, identification of co-fluctuation of functionally related genes, functional groupings of genes, and logical descriptions of gene regulation, among others. It is a starting point for understanding the large-scale network [13,44]. Domany [15] proposed a Coupled Two-Way Clustering (CTWC), which breaks down the total dataset into subsets of genes and samples that can reveal significant partitions into clusters. It provides clues about the function of genes and their roles in various pathologies.

The main contribution of this paper is a comprehensive and general purpose clustering approach that considers multiple objectives in the process and its application for clustering microarray data. The proposed approach has two components:

1. Multi-objective *K*-means Genetic Algorithm (MOKGA) based clustering approach has been developed to deliver a Pareto optimal clustering solution set without taking weight values into account. Otherwise, users need to consider several trials weighting with different values until a satisfactory result is obtained.
2. Cluster validity analysis and voting technique have been employed to evaluate the obtained candidate optimal number of clusters, by applying some of the well-known cluster validity techniques, namely Silhouette, C index, Dunn's index, DB index, SD index and S-Dbw index, to the clustering results obtained from MOKGA. It gives one or more options for the optimal number of clusters.

The applicability and effectiveness of the described clustering approach and clustering validity analysis process are demonstrated by conducting experiments using seven datasets from various domains: two breast cancer datasets, namely GSE12093 and GSE9195, Fig2data, NCI60 cancer data, Leukemia data sets available at Genomics Department of Stanford University, UCI machine learning repository, Iris at Genome Research MIT and Ruspini dataset.

The balance of the paper is organized as follows. Section 2 is an overview of the clustering approaches used primarily in the microarray data analysis area. Section 3 is devoted to the development of the new clustering system MOKGA for clustering both gene expression and general datasets. Section 4 reports experimental results on five datasets to test the applicability, performance, and efficiency of the system. Section 5 discusses the advantages and disadvantages of the proposed approach in comparison with other existing methods; conclusions are made and future research directions are suggested.

2. Related work

Existing clustering techniques which have been used for gene expression data can be classified into hierarchical clustering [28,48], partitioning [33], graph-based [44] and model-based [61,67] approaches.

Hierarchical clustering algorithms have been widely used in the area of gene expression data analysis. For example, Waddell and Kishino [67] applied hierarchical clustering based on partial correlations on NC160 gene expression data to find a tight and closed set of genes, and the interaction of important genes of the cell cycle. A tree structure *dendogram* is used to illustrate the hierarchical clustering [20,28,48]. Hierarchical clustering methods suffer from different aspects as stated by statisticians, including robustness, non-uniqueness, and inverse interpretation of the hierarchy [45,63]. Segal and Koller [59] proposed probabilistic abstraction hierarchies (PAH). This method improved the performance of traditional hierarchical clustering by handling the drawbacks mentioned above.

K-Means is a commonly used algorithm for partition clustering [33]. The purpose of *K*-Means clustering is the optimization of an objective function that is described by the equation:

$$E = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i) \tag{2.1}$$

where m_i is the center of cluster C_i, and $d(x, m_i)$ is the Euclidean distance between a point x and m_i. It can be seen that the criterion function attempts to minimize the distance between each point and the center of its cluster.

Self Organizing Maps (SOM) [30] is popular in vector quantization. It uses an incremental approach; points (patterns) are processed one-by-one. The shortcoming of SOM is that the size of the two dimensional grid and the number of nodes have to be predetermined. It suits well when prior information about data distribution is not available. Double self organizing maps (DSOM) technique [68] is also used for gene expression data clustering. In DSOM, each node does not have only an n-dimensional synaptic weight vector, but also a 2-dimensional position vector.

The model-based approach [53] is a promising technique, which assumes that data are generated by a mixture of finite number of probability distributions. In this approach, each cluster represents a probability distribution and a likelihood-based framework can be used. The Bayesian method is a model-based approach used in gene expression data analysis. Barash et al. [2,3] applied the Bayesian method on gene-expression time series data to study the response of human fibroblasts to serum. Gaussian mixture model is