Physica A 493 (2018) 162–176

Contents lists available at ScienceDirect

Physica A

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

Performance analysis of clustering techniques over microarray data: A case study

Rasmita Dash [a,*], Bijan Bihari Misra [b]

[a] Department of Computer Sc. & Information Technology, Institute of Technical Education and Research, Siksha'O' Anusandhan University, Khandagiri Square, Bhubaneswar, 751030 Odisha, India
[b] Department of Computer Sc. & Engineering, Silicon Institute of Technology, Bhubaneswar, 751024 Odisha, India

HIGHLIGHTS

- Stable clustering technique using a grading approach is suggested.
- A stable model (out of many) is evaluated irrespective of behavior of the dataset.
- Including hybrid swarm based approach, five clustering techniques are implemented.
- Significance of this approach is validated by Nemenyi post-hoc hypothetical test.

ARTICLE INFO

Article history:
Received 22 December 2016
Received in revised form 21 August 2017

Keywords:
Microarray data
Feature selection
Cluster analysis
Particle swarm optimization
Statistical test

ABSTRACT

Handling big data is one of the major issues in the field of statistical data analysis. In such investigation cluster analysis plays a vital role to deal with the large scale data. There are many clustering techniques with different cluster analysis approach. But which approach suits a particular dataset is difficult to predict. To deal with this problem a grading approach is introduced over many clustering techniques to identify a stable technique. But the grading approach depends on the characteristic of dataset as well as on the validity indices. So a two stage grading approach is implemented. In this study the grading approach is implemented over five clustering techniques like hybrid swarm based clustering (HSC), *k*-means, partitioning around medoids (PAM), vector quantization (VQ) and agglomerative nesting (AGNES). The experimentation is conducted over five microarray datasets with seven validity indices. The finding of grading approach that a cluster technique is significant is also established by Nemenyi post-hoc hypothetical test.

1. Introduction

Large size of scientific data brings a lot challenges in front of the researcher to recover useful information when traditional data mining techniques are used. Cluster analysis is one of the data mining approaches to deal with the large scale data. It is

* Corresponding author.
E-mail addresses: rasmitadash@soauniversity.ac.in (R. Dash), misrabijan@gmail.com (B.B. Misra).

https://doi.org/10.1016/j.physa.2017.10.032
0378-4371/

an unsupervised learning approach in which the entire data is divided into number of sub groups so that data with similar or related types forms a group called as a cluster. Here the cluster identification is efficient if the data in a cluster are more related with each other and data of different clusters are less related. Usually high inter cluster similarity and low intra cluster similarity measures are considered for better cluster quality.

But cluster analysis in the field big data analysis is a critical task. One of the real world high dimensional databases is gene expression data in Bioinformatics application. Microarray dataset or gene expression datasets are organized as matrix form and is experimented on different samples. The column represents different genes in gene expression data and row represents sample measured at different time point. Again in microarray data the number of genes is too large (in the range of 1000 to 10 000) and number of sample is comparatively small (in the range of 100). This, however, poses a great challenge to traditional clustering algorithms. So gene selection or feature selection is prerequisite for cluster analysis.

Feature selection is the process of identifying the most relevant feature from the dataset and representing the high dimensional data with a smaller space. But for microarray data, the most suitable feature selection determination is very difficult as the sample size is too small as compared to the number of genes. In the standardized microarray data some genes are highly correlated and out of hundred genes one gene is sufficient enough to describe the data. So considering one correlated feature is sufficient and the dataset is also reduced. In general the cluster quality is degraded substantially using all the features in such high dimensional data [1,2] So with feature selection both computation requirements and predictor accuracy can be improved. Hence for cluster analysis, feature selection is prerequisite and relevant gene selection is important.

There are many clustering techniques for high dimensional data analysis. These are categorized into partition based clustering, hierarchical clustering and density based clustering. But these clustering techniques have different cluster analysis approach. But which approach suits a particular dataset is difficult to predict. It depends on both features of the dataset as well as on the validity indices. So it is required to identify a suitable (or stable) clustering approach for high dimensional data analysis. In this research we have presented a case study through which the above mentioned issues are handled.

In the initial stage of proposed work statistical measures are implemented to discard the insignificant features. Feature with similar value do not take participation to identify a cluster. So these features are discarded and dissimilar value features are kept in the reduced feature set. So standard deviation of all the features is evaluated. Then a threshold value on standard deviation is used to discard the irrelevant features. Then cluster analysis is done using five cluster technique like HSC, *k*-means, PAM, VQ and AGNES. The experiment is conducted taking five microarray datasets and the result is shown using seven clustering indices. The performance of a cluster approach depends on both datasets and cluster validity indices so a grading technique is introduced to justify our approach. In the 1st stage grading of cluster approach is done with respect to datasets and in the second stage with respect to validity indices. Finally implementing a statistical non parametric approach i.e. Nemenyi post hoc test, our result is validated.

Rest part of the paper is organized as follows. A literature review on clustering approaches is presented in Section 2. In Section 3 the strategy for the case study has been discussed. The detail descriptions of proposed clustering approach along with other clustering techniques are highlighted in Section 4. The experimental set up and result analysis is presented in Section 5. It includes the dataset used for the experimental analysis, data normalization process, validity indices and result analysis. Finally the concluding remark is highlighted with Section 6.

2. Literature survey

One of the important gene expression data visualization technique is an unsupervised learning approach called as cluster analysis. Basically cluster analysis is used to extract more correlated genes in microarray experimentation. So genes with in a group shows similar behavior in and genes of two different groups are more or less dissimilar their expression levels.

There are many kinds of traditional clustering approaches used and adopted in microarray environment. Some clustering approaches are proposed in these literatures [3–6] especially for microarray data. Further many biclustering algorithms are introduced for microarray data which can be found in the literature. The first biclustering algorithm is proposed by Cheng and Church [7]. These techniques are tested on previously available gene datasets. Authors in [8,9] used *K*-means algorithms for gene data clustering. Currently many variants of *K*-means algorithms are used for data clustering are in these literatures [9–11]. Few other algorithms are also implemented to overcome the drawbacks of *K*-means based clustering [12,13]. Self organizing maps (SOM) are also used for cluster analysis of microarray data and even in some cases in shows better performance as compared to Kmeans algorithm [12]. Chang et al. proposed [7] a SOM based clustering approach for extracting biological information from gene expression data. In this contribution a query based SOM (QBSOM) is compared with simple SOM for gene expression data analysis. Finding shows that the computational cost of simple SOM is more than 65% greater than QBSOM.