I J C T A, 10(9), 2017, pp. 353-360
ISSN: 0974-5572

A comparative Study of Clustering Algorithms for Cancer Data Analysis

Ajaya Kumar Parida*, Tapas Ranjan Baitharu and Subhendu Kumar Pani*****

ABSTRACT

Clustering is the most general form of automatic unsupervised data analysis. Traditionally, clustering techniques work by trying to find relationships in the data forming groups (clusters) using only the information present in the data, aiming to fulfil two goals: maximise the similarity between the data points which are assigned to the same cluster and keep the data points assigned to different clusters as dissimilar as possible. Research efforts dedicated to data mining, which focussed on data clustering, have recently been undergoing a tremendous change. In this work, effort has been made to compare between few Clustering algorithms such as: K means, Hierarchical, DBSCAN, COBWEB, FARTHEST FIRST ALGORITHM using cancer dataset.

Keywords: K means, Hierarchical, DBSCAN, COBWEB, FARTHEST FIRST ALGORITHM

1. INTRODUCTION

CLUSTERING is a data mining method to group the like data into a cluster and dissimilar data into different clusters. Clustering can be regarded as the most important unsupervised learning technique so as every other problem of this kind, it deals with finding a structure in a set of unlabeled data. Clustering is ¯the method of organizing objects into groups whose members are similar in some way. A cluster is therefore a group of objects which are ¯similar between them and are ¯dissimilar to the objects belonging to other clusters. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters).

Data mining as a multidisciplinary joint effort from databases, machine learning, and statistics, is championing in turning mountains of data into nuggets [1, 2, 3]. Researchers and practitioners realize that in order to use data mining tools effectively, data pre-processing is essential to successful data mining..

Data clustering is a method of putting similar data into groups. A clustering technique partitions a data set into several groups such that the similarity within a group is larger than among groups. Furthermore, the majority of the data collected in many problems seem to have some intrinsic properties that lend themselves to natural groupings. Clustering techniques are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. Finding these groupings or trying to categorize the data is not a simple task for or three dimensions at maximum.) Another reason for clustering is to find out relevance knowledge in data. Data cluster are created to meet particular requirements that cannot created using any of the categorical levels. One can merge data subjects as a temporary group to get a data cluster[10, 11, 12].

Knowledge discovery method consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation.

* Associate Professor, Dept. of IT, OEC, BPUT, Odisha, India, *Email: erakparida@gmail.com*
** Associate Professor, Dept. of CSE, OEC, BPUT, Odisha, India, *Email: trbaitharu2001@yahoo.co.in*
*** Associate Professor, Dept. of CSE, OEC, BPUT, Odisha, India, *Email: skpani.india@gmail.com*
Corresponding Author: Dr. Subhendu Kumar Pani

Data mining functionalities are characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis. Three of the main data mining techniques are regression, classification and clustering.

2. CLUSTERING TECHNIQUES

Dividing objects in meaningful groups of objects or classes (cluster) based on common characteristic, play an important role in how people analyse and describe the world. For an example, even children can quickly label the object in a photograph, such as buildings, trees, people and so on.

2.1. Cluster Analysis

Cluster Analysis technique as a field grew very quickly with the goal of grouping data objects, based on information found in data and describing the relationships inside the data. The purpose is to separate the objects into groups, with the objects related (similar) together and unrelated with another group of objects. It is being applied in variety of science disciplines and has been studied in myriad of expert research communities such as machine learning, statistic, optimization and computational geometry [5, 6].

The following are some examples:

- *Biology*: Biologists when they a long time ago created a taxonomy (hierarchical classification) made a form of clustering according to genus, family, species and so on.

  But also recently they have applied clustering to analyse the myriad amount of genetic information, such as a group of genes that has similar functions[7].
- *Information Retrieval*: The World Wide Web consists of billions of web pages that are accessed with the help search engine queries. Clustering can be used to create small clusters of search results.[8]
- *Psychology and Medicine*: Clustering techniques are used to analyse frequent conditions of an illness and identifying different subcategories. For example, clustering is used to identify different types of depression, and cluster analysis is used to detect patterns in the distribution/spread of a disease.
- *Business*: In this field there exists a large amount of information on current and potential customers. Clustering helps to group customer activities, as previously mentioned in detail.

Assume we have twelve points and three different ways to dividing them into clusters.

2.2. Clustering Methods

Clustering methods can be classified into the following categories

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

2.2.1. Partitioning Method

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember:

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

2.2.2. Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here "

- Agglomerative Approach
- Divisive Approach

AGGLOMERATIVE APPROACH

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

DIVISIVE APPROACH

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone[9].

APPROACHES TO IMPROVE QUALITY OF HIERARCHICAL CLUSTERING

Here are the two approaches that are used to improve the quality of hierarchical clustering—

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

2.2.3. Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

2.2.4. Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

2.2.5. Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

2.2.6. Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

3. COMPARISION AND RESULT ANALYSIS

3.1. WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

3.2. Results Analysis

We run clustering algorithms on the original dataset to generate scenarios for comparative analysis of the clusters. Figure 1 shows Visualize Cluster Assignments Using Simple K-Means Clustering Algorithm. Figure 2 shows Visualize Cluster Assignments Using Density Based Clusterer Algorithm. Figure 3 shows Visualize Cluster Assignments Using Hierarchical Clusterer Algorithm. Figure 4 shows Visualize Cluster Assignments Using Simple FarthestFirst Algorithm Table 1 shows the comparative performance of the clustering algorithms. We performed computer simulation on a Lung Cancer dataset dataset available UCI Machine Learning Repository [4].

3.3. Cluster Assignments

Recently data mining techniques have encompassed every field in our life. Data mining techniques are being used in the medical, banking, insurances, education, retail industry etc. Prior to working in the data

Table 1
The comparative performance of the clustering algorithms

Name	*Data set name*	*No. of clusters*	*Cluster distribution*	*No of iterations*	*Sum of squared error*	*Log like-lihood*	*Time taken to build model (sec)*
K-Means	Lung Cancer	02	67(33%) 136(67%)	09	51575.3602		10.88
Density Based Clusterer	Lung Cancer	02	70(34%) 133(66%)	09	51575.3602	-61729.70	15.13
Hierarchical Clusterer	Lung Cancer	02	20201				42.0
Farthest First	Lung Cancer	02	191(94%) 12(6%)				7.22

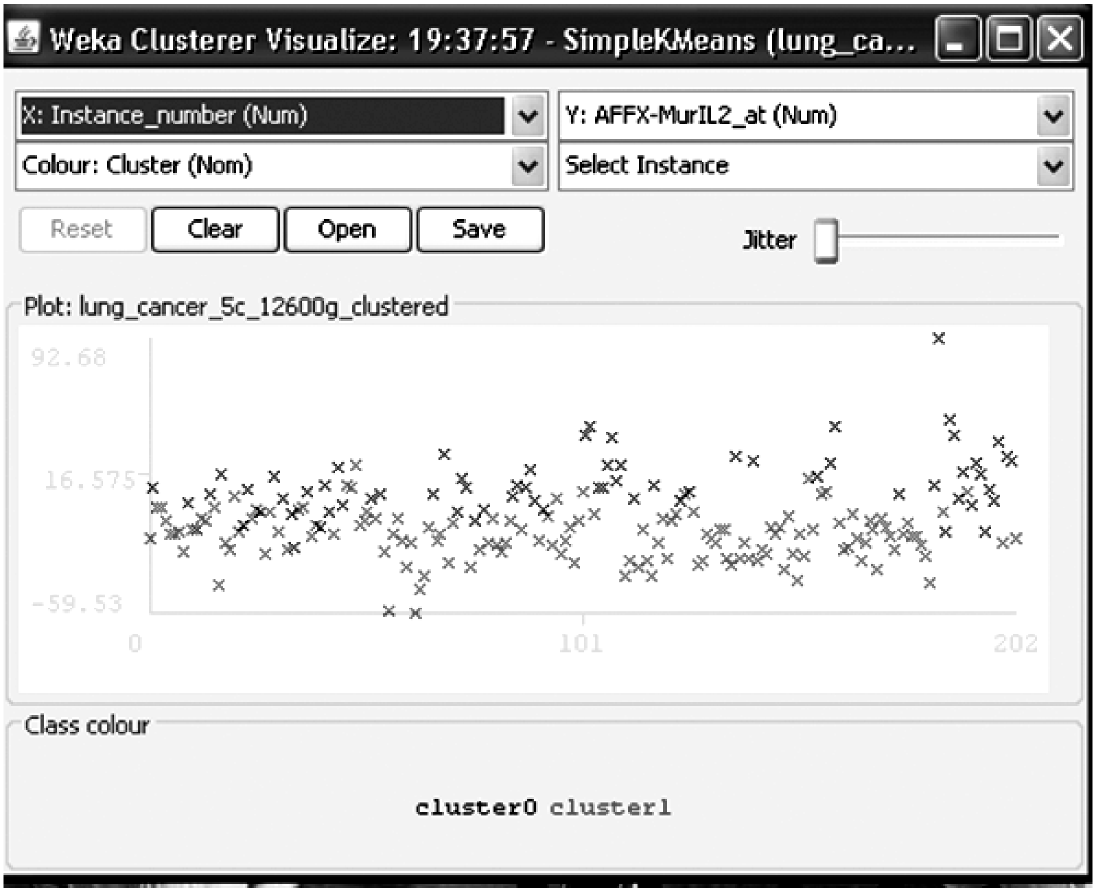

Figure 1: Visualize Cluster Assignments Using Simple K-Means Clustering Algorithm

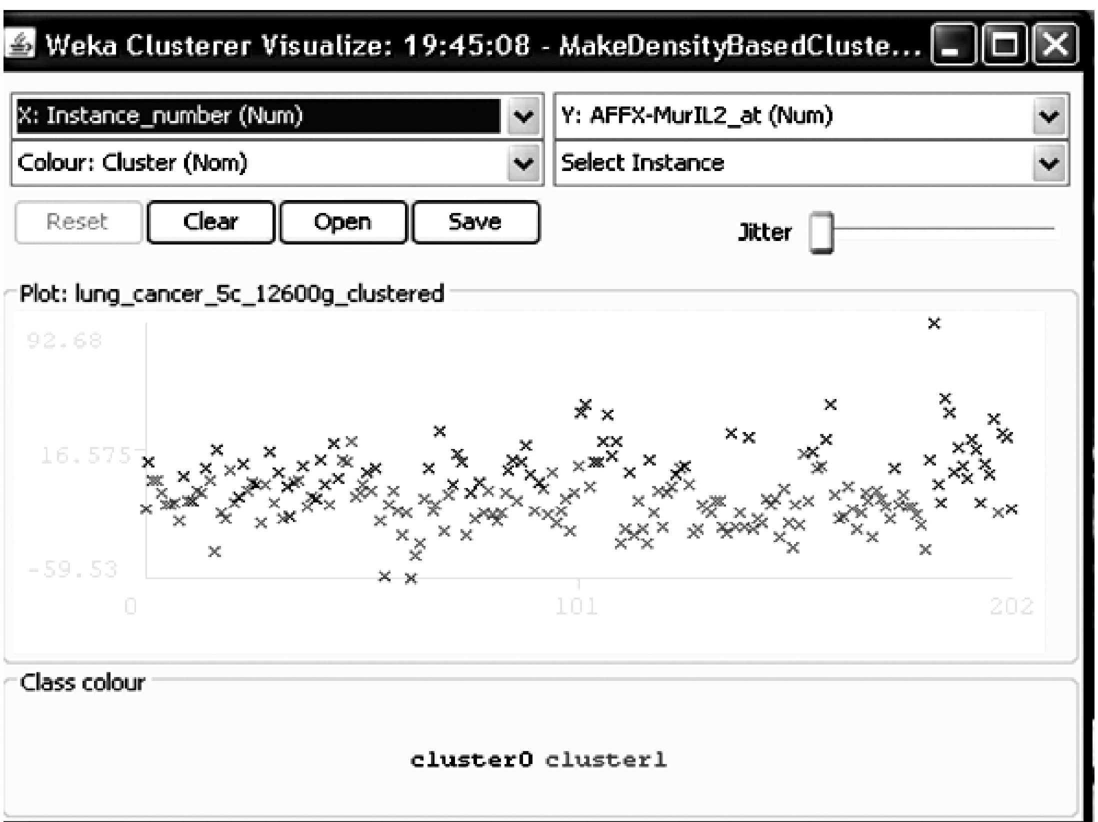

Figure 2: Visualize Cluster Assignments Using Make Density Based Clustering Algorithm

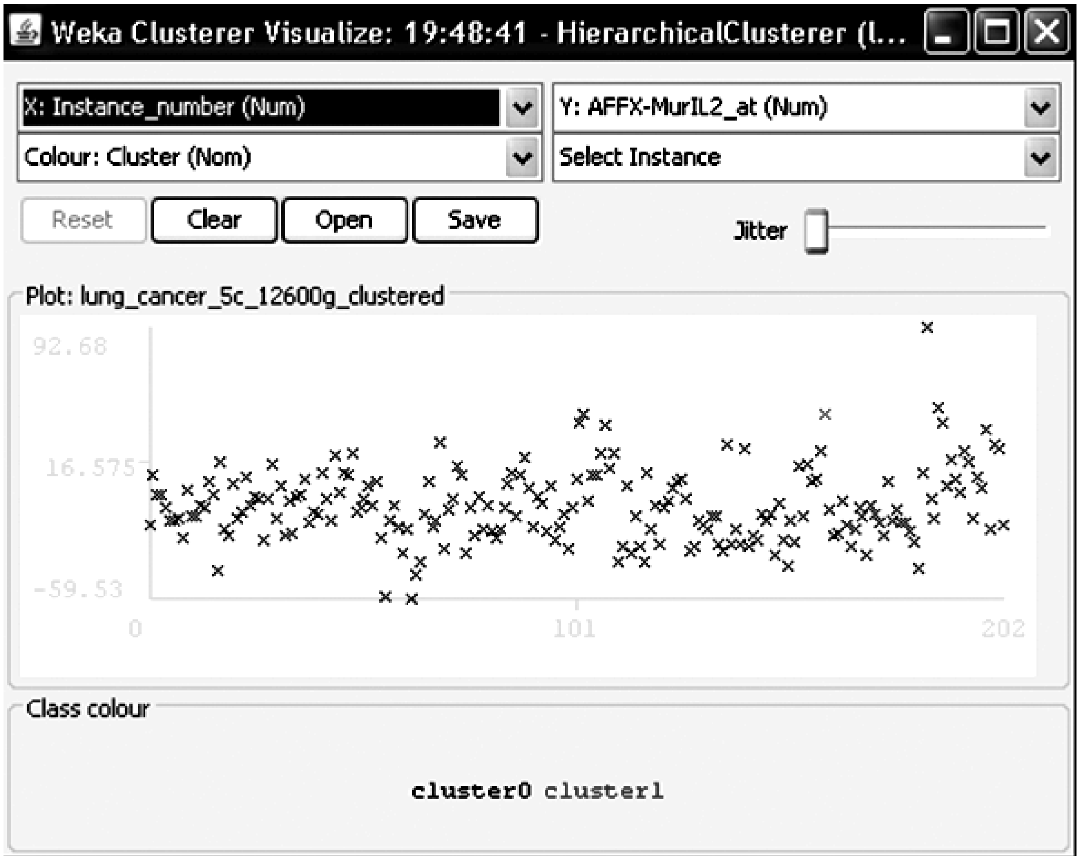

Figure 3: Visualize Cluster Assignments Using Hierarchical Clustering Algorithm

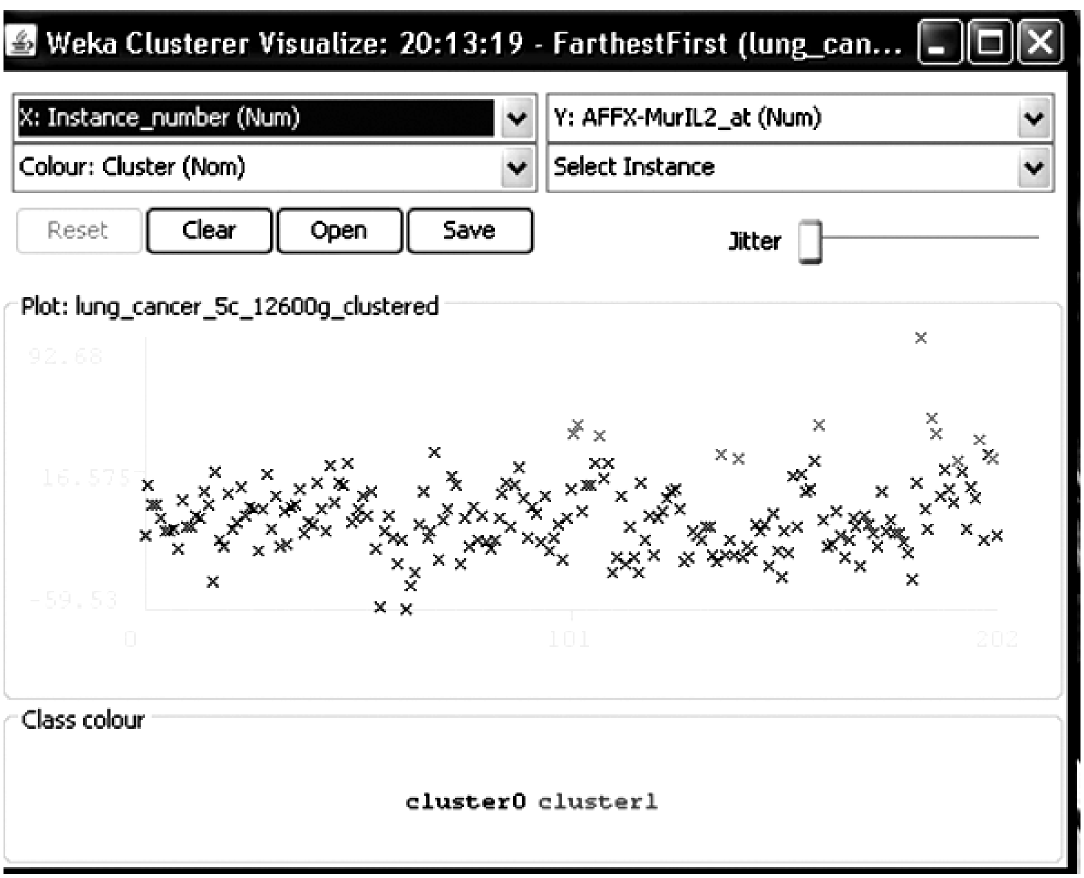

Figure 4: Visualize Cluster Assignments Using Farthest First Clustering Algorithm

mining models, it is very important to have the knowledge of the existing essential algorithms.[14] Every algorithm has their own significance and we use them on the nature of the data, but on the basis of this research we concluded that k-means clustering algorithm is simplest algorithm as compared to other algorithms and its performance is better than Hierarchical Clustering algorithm. Density based clustering algorithm is not suitable for data having very huge variations in density and hierarchical clustering algorithm is more susceptible to noisy data. EM algorithm takes more time to build cluster as compared to K- Mean, hierarchical, density based clustering algorithms, that's why k-mean and density based algorithm are better than EM algorithm. Density based algorithm takes relatively less time to build a cluster but it's not better than the k-mean algorithm since density based algorithm has high log likelihood value, if the value of log likelihood is high then it makes bad cluster. Hence k-mean is best algorithm because it takes very less time to build a model. Hierarchal algorithm take more time than k-mean algorithm and cluster instances are also not good in hierarchal algorithm.

Clustering is a vivid method. The solution is not exclusive and it firmly depends upon the analysts' choices. Clustering always provides groups or clusters, even if there is no predefined structure. While applying cluster analysis we are contemplating that the groups exist. But this speculation may be false. The outcome of clustering should never be generalized.[13]

4. CONCLUSION

we conducted an experiment to find the predictive performance of different clustering techniques. We select 4 popular clustering techniques considering their qualitative performance for the experiment. We formulate different Cluster Assignments for a given dataset drawn from Lung Cancer available at UCI machine learning repository. After analysing the quantitative data generated from the computer simulations, we find that the k-mean is best algorithm because it takes very less time to build a model. we have made an attempt to identify the problems associated with clustering of gene expression data, using traditional clustering methods, mainly due to the high dimensionality of the data involved. For this reason, subspace clustering techniques can be used to uncover the complex relationships found in data since they evaluate features only on a subset of the data. Differentiating between the nearest and the farthest neighbours becomes extremely difficult in high dimensional data spaces.

5. FUTURE WORK

In this paper, we analysed the performance of most popular clustering algorithms. The experimental data show mixed results. We propose to extend our work by considering multiple datasets drawn from different domains, so that the results will be sound enough for generalization. In this paper we have intended to compare the pre-defined four algorithms and we have given some conclusions above. But still we were not able to cover all the factors for comparing these four algorithms. As a future work, comparison between these algorithms (or may other algorithms) may be done using different parameters other than considered in this paper.

REFERENCES

[1] Han J. and Kamber M., Data mining concept and techniques, Morgan Kaufmann Publishers, London, 2001.

[2] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[3] Liu, H. Feature Extraction, Construction and Selection: A Data Mining Perspective, ISBN0-7923-8196-3, Kluwer Academic Publishers, 1998.

[4] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german.

[5] Sara C. Madeira and Arlindo L. Oliveira. A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. Algorithms for Molecular Biology, 4(8), June 2009.

[6] W. Ayadi and M. Elloumi. Algorithms in Computational Molecular Biology : Techniques,Approaches and Applications. chapter Biclustering of Microarray Data, 2011.

[7] Sara C. Madeira and Arlindo L. Oliveira. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 1:24–45, 2004. ISSN 1545-5963.

[8] Law Ngai-Fong Siu Wan-Chi Cheng, Kin-On and Alan Wee-Chung. Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. BMC Bioinformatics, 2008.

[9] Xiaowen Liu and Lusheng Wang. Computing the maximum similarity bi-clusters of gene expression data. Bioinformatics, 23(1):50–56, 2007.

[10] Aguilar-Ruiz and Jesús S. Shifting and scaling patterns from gene expression data. Bioinformatics, 21(20): 3840–3845, 2005.

[11] Daniel Gusenleitner, Eleanor Howe, Stefan Bentink, John Quackenbush, and Aedin C. Culhane. ibbig: iterative binary bi-clustering of gene sets. Bioinformatics, 28(19):2484–2492, 2012.

[12] Lazzeroni and Owen. Plaid models for gene expression data. Statistica Sinica., 2002.

[13] Shawn Mankad and George Michailidis. Biclustering three-dimensional data arrays with plaid models. Journal of Computational and Graphical Statistics, 2013.

[14] Ole Andreatta, Massimo Lund and Morten Nielsen. Simultaneous alignment and clustering of peptide data using a gibbs sampling approach. Bioinformatics, 29(1):8–14, 2013.