بسم الله الرحمن الرحيم

∞∞∞

تم رفع هذه الرسالة بواسطة / سامية زكى يوسف

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى مسئولية عن محتوى هذه الرسالة.

ملاحظات: لا يوجد

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Gene Signatures Prediction of Genetic Diseases

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences

By

Hassan Sayed Ramadan Barakat

B.Sc. of Computer & Information Sciences,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Dr. Khaled El-Bahnasy
Professor in Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Prof. Dr. Mohamed El-Eliemy
Professor in Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Huda Amin Maghawry
Lecturer in Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Cairo 2022

Acknowledgment

We thank Dr. Feryal Farouk Sherif (National Research Centre) for providing the clinical criteria used for dividing breast cancer dataset.

Abstract

Over the last few years, several standard clustering approaches have been proposed to evaluate gene expression data. On the other hand, identifying breast cancer subtypes with consistency is difficult.

DBSCAN-BICLIC, a modified BICLIC biclustering algorithm, was proposed to discover signature genes of breast cancer subtypes. DBSCAN-BICLIC gives an efficient solution for properly clustering seeds. Experimental results on 2509 breast cancer situations were evaluated using clinical data. The analysis resulted in the division of breast cancer conditions into 22 groups, each with its own set of clinical data.

Clinical criteria that are similar will be gathered together. DBSCAN-BICLIC discovered the biomarkers for each group of conditions, which is an intriguing part of the experiment.

As a result, each subtype of breast cancer has its own set of signature genes. For 10 groups, the DBSCAN-BICLIC algorithm was used. The top five effective signature genes that have been created for each group are presented. DBSCAN-BICLIC has identified 40 signature genes across all 10 groups. According to the literature, 32 of them have been validated as signature genes. The 32 genes are extremely effective breast cancer prognostic genes.

Although the promising results of DBSCAN-BICLIC, but it cannot work automatically with any biological dataset. This is because of the epsilon parameter of DBSCAN clustering algorithm, and most of researchers choose it randomly. Therefore, the objective was to propose a heuristic approach to find the optimal epsilon for DBSCAN clustering algorithm. The concept of this approach is to repeat DBSCAN many times, and each time it calculates a different epsilon value till it finds the optimal epsilon. Finding the optimal epsilon depends on evaluating clusters each time, and for sure optimal epsilon has the best evaluation scores. Proposed approach uses the root mean square standard deviation (RMSSTD), and the R-squared (RS) to evaluate clusters. The proposed approach had been run on three benchmark different dimensional datasets. Also, Silhouette index was used to validate the clustering results of the proposed approach. The proposed approach was successfully able to find the optimal epsilon for all three datasets.

List of Publications

1. H. S. Ramadan, H. A. Maghawry, and K. El-Bahnasy, “Determination of signature genes of breast cancer subtypes evaluated by clinical criteria using biclustering algorithm,” 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), pp. 431–441, 2021.
2. H. S. Ramadan, H. A. Maghawry, M. El-Eleamy, and K. El-Bahnasy, “A Heuristic Novel Approach for Determination of Optimal Epsilon for DBSCAN Clustering Algorithm,” Journal of Theoretical and Applied Information Technology (JATIT), vol. 100, no. 7, pp. 2243–2253, 2022.
3. H. S. Ramadan, and K. El-Bahnasy, “A review of clustering algorithms for determination of cancer signatures,” International Journal of Intelligent Computing and Information Sciences (IJICIS), vol. 22, no. 3, pp. 138–151, 2022.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
BICLIC	Biclustering by Correlated and Large number of Individual Clustered seeds
BR	Breast Cancer
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCIS	Ductal carcinoma in situ
Eps	Epsilon
ER	Estrogen Receptor
HER2	Human Epidermal Growth Factor Receptor 2
LCIS	Lobular carcinoma in situ
MinPts	Minimum Points
mRNA	Messenger Ribonucleic Acid
PR	Progesterone Receptor
RMSSTD	Root Mean Square Standard Deviation
RS	R-squared

Chapter 1

Introduction

Chapter 1. Introduction

1.1 Overview

Breast cancer is a type of cancer that develops in the breast cells. It is a disease that is both complex and heterogeneous.

Breast cancer is the second most common cancer diagnosed in women in the United States, after skin cancer. It can strike both men and women, although it affects women significantly more frequently. It is one of the leading causes of cancer-related death among women.

There are several types of breast cancer. The type of breast cancer you have is determined by where it started in the breast and other factors. So, because of these types, and subtypes, it’s hard to develop a generic treatment for all breast cancer patients.

Also, over time, the prognosis of breast cancer patients has been improved. However, only portion of the patients that current therapy has effect on. Further advancements in tailored treatment for breast cancer patients are likely to solve this challenge.

The definition of breast cancer molecular subtypes has been defined thanks to gene expression profiling. It's a big step forward in the right direction. Most studies use a published 'intrinsic gene list' [1] to undertake gene expression analysis. Breast cancer has lately been classified into subgroups based on expression patterns.

Several methodologies [2, 3], such as hierarchical cluster analysis, can be used to analyze patterns in gene expression data.

Hierarchical clustering groups conditions based on expression similarity across all genes. Only when looking for global patterns are traditional clustering techniques successful.

They only affect a small number of genes and/or situations, although there are numerous regulatory patterns.

Bi-clustering algorithms [4, 5] have been developed to find local patterns in biological data [6, 7]. A bicluster is a subgroup of genes that are co-expressed in just a small number of samples. Bi-clustering algorithms are an efficient way to find the signature genes of breast cancer. Biclustering by Correlated and Large number of Individual Clustered seeds (BICLIC) is a biclustering algorithm [8]. However, BICLIC biclustering results might be not accurate because clustered seeds are clustered using standard deviation, which is not a clustering algorithm.

So, a hybrid approach titled DBSCAN-BICLIC (based on DBSCAN), is proposed to identify signature genes of breast cancer subtypes, after classifying breast cancer into subtypes according to specific clinical criteria. DBSCAN-BICLIC embedded the result of DBSCAN clustering in BICLIC seed clustering as prior knowledge. As noticed the hybrid approach depends on DBSCAN clustering algorithm, and one of the parameters of DBSCAN is the epsilon (Eps), which defines the radius of neighborhood around a point x. Most of researchers choose it randomly. Therefore, heuristic approach is proposed to find the optimal epsilon for DBSCAN clustering algorithm.

1.2 Motivation

Biclustering is a powerful data mining technique that allows clustering of rows and columns, simultaneously, in a matrix-format data set. It has become a common method for analyzing gene expression data, particularly for identifying functionally related gene sets under various experimental settings. The quality of biclusters is usually determined by a metric or cost function in most biclustering systems.

Biclustering algorithms are efficient, and reliable in finding the local patterns in gene expression data, which is suitable to identify the biomarkers.

1.3 Objectives

Most of mRNA gene expressions datasets have the genes of the hole human genome, so, there are many unrelated breast cancer genes. One of the objectives is to filter only the related genes of breast cancer.

The are many subtypes of breast cancer. To identify the signature genes for all conditions accurately, the conditions must be divided into groups, and each group is considered as a subtype of breast cancer. Therefore, original clinical criteria will be as a followed to divide breast cancer patients into groups.

To identify signature genes of each breast cancer groups, a hybrid approach titled DBSCAN-BICLIC will be proposed. It will improve the prognosis of breast cancer patients efficiently. The hybrid approach cannot be for any biological dataset, because it has a parameter, which is required to be set manually. This parameter is the epsilon of DBSCAN clustering algorithm.

So, a heuristic novel approach will be proposed to determine the optimal epsilon for DBSCAN clustering algorithm accurately.

1.4 Methodology

It starts with selecting genes that related to breast cancer. ‘Intrinsic gene list’ [1], and Ion RNA AmpliSeq Kits were used to extract the only related-genes of breast cancer.

Clinical sheets of dataset used in study of Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016) [9], [10], [11], [12] explains how conditions are different, and cannot be in one group. They have various mRNA gene expressions, which will cause inaccurate results. To get accurate results, each group should have similar conditions, and the similarity here is on the basis of clinical information. The clinical criteria depend on specific fields of the dataset sheet attached to the previously mentioned dataset.

These clinical criteria will be used to divide dataset into groups of conditions.

DBSCAN-BICLIC will be run for each group will be generated by the previously mentioned clinical criteria. DBSCAN-BICLIC will identify signature genes of breast cancer subtypes.

DBSCAN clustering algorithm used in the hybrid approach DBSCAN-BICLIC requires Eps value as a parameter. To find the optimal value of Eps used, genes of the mentioned dataset have been clustered several times for each group of conditions.

This way is not efficient, and not reliable to be used every time with different datasets. So, a heuristic novel approach is proposed to determine the optimal Eps of DBSCAN clustering algorithm.