Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765934

AN EXTENSIVE STUDY ON CARDIOMYOPATHY CLASSIFICATION TECHNIQUE USING MICROARRAY DATA

T.Sangeetha
Research Scholar, Department of Computer Science, PSG College of Arts & Science
Coimbatore - 641014, Tamil Nadu, India

Dr.K.Manikandan
Associate Professor, Department of Computer Science, PSG College of Arts & Science
Coimbatore - 641014, Tamil Nadu, India

Abstract

Cardiomyopathy is one of important cause of chronic heart failure which makes heart muscle harder to pump blood to other part of the body which leads to high mortality rate. Hence it is becomes mandatory to diagnosis and predict the disease in order to prevent the person against heart failure. However manual analysis of the disease is highly complex and leads to poor prognosis. In order to alleviate those challenges and predict the disease in early stage, many risk assessment methods has been modeled using machine learning and deep learning paradigms using genome wide association studies. Especially Cardiomyopathy risk assessment through gene expression from microarray data provides excellent results. In this article, various architectures to identify the Cardiomyopathy on gene expression profiles of the GEO databases has been analyzed. Initially gene expression profile is processed using normalization technique to regularize the down regulated and unregulated genes in specified range. Next feature extraction technique to obtain the differentially expressed gene. FurtherPotential biomarker is employed to select the DCM related genes such as MYH6, PTH1R, ADAM15, S100A4CKM, NKX2–5 and ATP2A2 which contains the mutated chromosomes. Finally classifier model is employed to the discriminate the core set of genes with core set of target genes extracted from the diseased patient of the mutated chromosomes related to Cardiomyopathy which is considered as ground truth data. Experimental analysis of various classifier employed to the classify the core set of genes into type of classes of Cardiomyopathy is carried out on interfering the results of the classifier on the cross fold validation. Performance evaluation of the architectures on the mentioned dataset is performed using performance measure.

Keywords: Cardiomyopathy, Classification, Microarray data, Target Genes, Gene Profiling, Normalization, mRNA

1. Introduction

Cardiomyopathy is leading cause of chronic heart failure which increases the risk of prognosis of the patient. Hence it is becomes mandatory to diagnosis and predict the disease earlier in order to avoid the adverse effect of the disease. Cardiomyopathy is heart muscle diseases which experienced in the left ventricles and leading to systolic dysfunction and contractile function of the ventricle. Further it leads to poor blood circulation around the body. In order to alleviate those challenges and predict the disease in early stage , many risk assessment methods has been modeled using machine learning and deep learning paradigms using genome wide association

studies. Especially Cardiomyopathy risk assessment through gene expression from microarray data provides excellent results.

In this article, various architectures to identify the Cardiomyopathy on gene expression profiles of the GEO databases have been analyzed. Initially gene expression profile is processed using normalization technique to regularize the down regulated and unregulated genes in specified range. Next feature extraction technique to obtain the differentially expressed gene. Further Potential biomarker is employed to select the DCM related genes such as MYH6, PTH1R, ADAM15, S100A4CKM, NKX2–5 and ATP2A2 which contains the mutated chromosomes. Finally classifier model is employed to the discriminate the core set of genes with core set of target genes extracted from the diseased patient of the mutated chromosomes related to Cardiomyopathy which is considered as ground truth data.

Rest of the article organized as follows, section 2 defines preliminary details of Cardiomyopathy disease, microarray gene expression dataset and RNA information of the gene expression data. section 3 details the problem statement. Section 4 details the analysis of the feature extraction and feature selection techniques to gene expression data related to Cardiomyopathy disease ,analysis of classification architecture to classify the disease on the core set of genes and its results on various performance measures. Finally article is concluded in the section 5

2. Preliminaries

In this section, preliminary details of the disease, data acquiring technologies employed to acquire the gene expression data for disease classification and processing elements of the gene expression data has been defined**.**

2.1. RNA

Functional RNA is single stranded molecule which is composed of shorter chain of nucleotides. It contain phosphate groups and sugar ribose. RNA carries the genetic information to cell as protein. Major type of RNA molecule is messenger RNA. RNA has chain of chemicals known as base which is termed as Adenine, Cytosine, Guanine and Thymine[1].

2.2. Genome

Genome is complete set of nucleic acid sequence of RNA which consists of the 23 pairs of chromosome located in the cell. It is represented as genetic information of an organism. It composed of information need to build the organism. It combines structures in biological process and molecular functions with potential molecular signatures.

2.3. Gene

Gene consist of set of function or instruction which controls thefunctional RNA on the chromosomes in the nucleus of the cell and it acts as instruction to make protein synthesis as it is building block of muscles, connecting tissues and skin. Gene comes in pairs. Changes in gene mutation cause's irregularities in making a protein. The irregular protein leads to genetic disorder[2].

3. Problem Statement

Manual Prediction of the Cardiomyopathy on the symptom of the patient leads to several complication in patients due to wrong treatment and it may cause to life threatening. Further it is mandatory to handle the disease using clinical data such as microarray gene

expression data. Risk Assessment method employed to process the gene expression data leads to discordant results due to follow reason

- Genetic susceptibility due to inconsistent gene expression pattern on protein transport and cellular catabolic process
- Large number of disease associated genes is processed to classify the disease and it fails to select the optimal genes
- It fails in discriminating the major forms of Cardiomyopathy disease stages.

4. **Analysis of the Microarray Gene expression data**

Gene expression data is the process in which information encode in gene is turned to function to present thousand of cells which is correlated with the corresponding protein. It is analyzed and interpreted using methods

- **Gene set enrichment analysis**

It is to provide functional annotation to the differentially expressed genes. It helps to identify certain biological process or molecular function of the nuclei of the cell. It is used to analyze both down regulated and up regulated genes. Data is represented in the array[3]. It is considered as finite structure analysis which fails to capture indirect association of the genes on the cellular processes in the biological process.

- **Heat Maps**

Heat map is to analyzing differentially expressed genes on basis of grouping the genes together based on the similarity of the gene expression pattern[4]. It is capable to identify genes which are regulated and associated with particular condition. Data is represented in the grid as it contains the reduced dimensions of genes. Further heat map has low sensitivity and specificity.

5. Analysis of Feature extraction technique

Feature extraction is to extract the differentially expressed genes after normalization of down and up regularized genes. Feature extraction technique to microarray data is as follows

5.1. Principle Component analysis

Principle component analysis was performed on the 2000 most variable genes across samples. It identifies any statistically significant genes which are considered as highly interconnected genes. Gene with large variance across samples is considered as principle component[5]. Variance is measure on the gene represented in covariance matrix. Eigen value represents the principle gene for further processing. Figure 1 represents the distribution of the gene with principle variance.

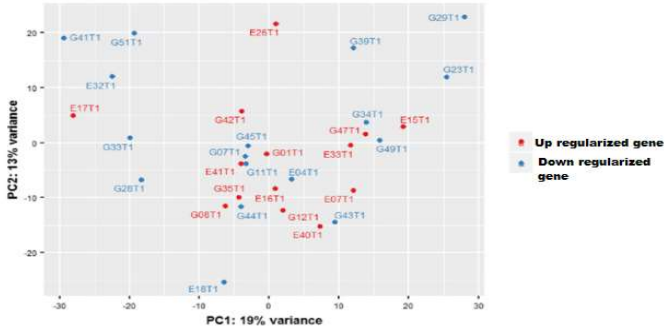

Figure 1: Gene distribution

5.2. Linear Discriminant Analysis

Linear discriminant analysis is used to compute the linear gene and non linear gene on the gene expression data on mean and standard deviation measurement through scatter matrix. Scatter matrix contains the non linear genes. Figure 2 represents the linear regularization of the genes[6]

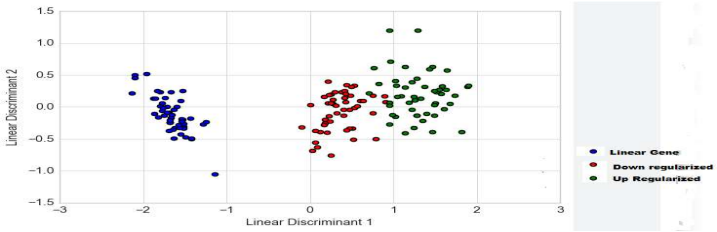

Figure 2: Regularization of gene

6. Analysis of feature selection technique

Feature selection is to select the optimal genes for effective classification of the disease on basis of its type based on fitness function. Feature selection technique to microarray data to compute core set of feature is as follows

6.1. Genetic Algorithm

Genetic algorithm is to select the gene using biomarkers. Core set of feature is selected on basis of fitness function on the operation of cross over and mutation operation of the chromosomes on the available feature extracted which represents the gene[7]. Figure 3 represent the gene selection of the extracted genes. It selects MYH6, PTH1R, ADAM15, S100A4CKM, NKX2–5 and ATP2A2 which contains the mutated chromosomes.

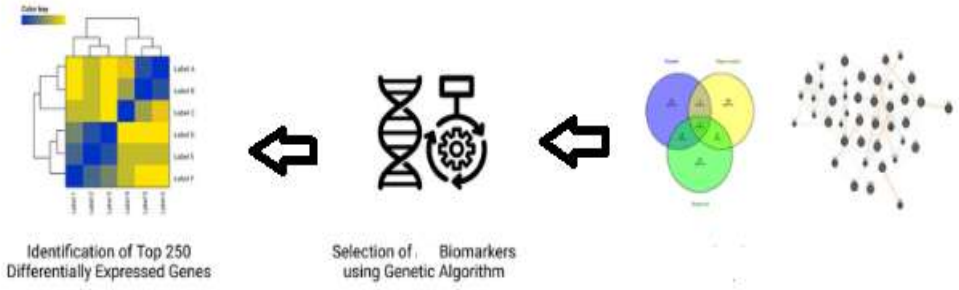

Figure 3: Feature selection using genetic algorithm

6.2. Particle Swarm Optimization

Particle Swarm optimization predict the optimal gene using fitness function. optimal gene is selected on the operation of the global best and local best gene for the Cardiomyopathy related disease. In this extracted gene is considered as particle and large muted gene is represented as velocity[8]. Figure 4 represents the gene selection using pbest and gbest.

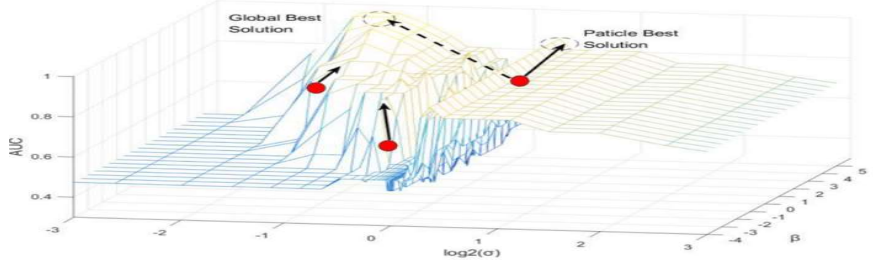

Figure 4: Feature Selection using PSO

7. Analysis of Classification Techniques

Feature Classification on the core set of features is to classify the disease on basis of its type using both machine learning and deep learning architectures. Classification architecture to microarray data is as follows

7.1. Support Vector Machine

Support Vector Machine classifies the Cardiomyopathy into ischemic Cardiomyopathy ,dilated Cardiomyopathy and neurofibromatosis. It uses the hyper plane and margin to classify the core set gene into the classes of the disease with support vector. It is machine learning classifier with decision boundaries[9]. Figure 5 represents the support vector machine classifier for Cardiomyopathy disease classification.

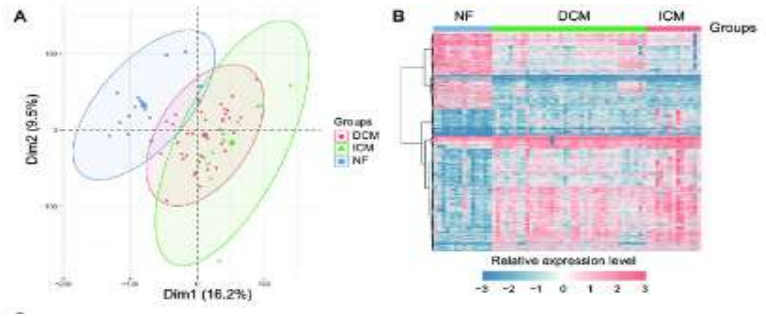

Figure 5: Support Vector Machine Classifier

7.2. Convolution Neural Network

Convolution Neural Network[10] classifies the Cardiomyopathy into ischemic Cardiomyopathy ,dilated Cardiomyopathy and neurofibromatosis. Figure 6 represents the architecture of Convolution Neural Network.

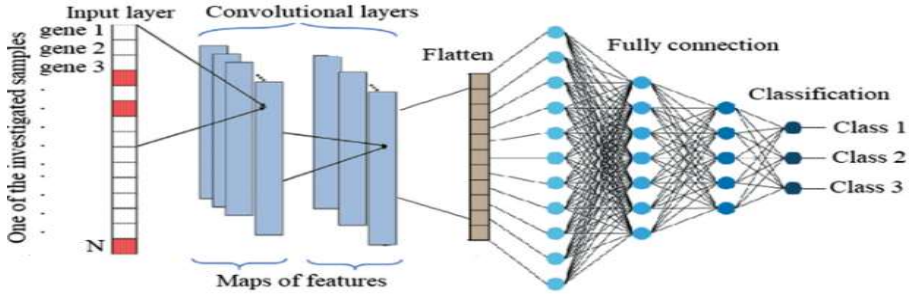

Figure 6: Feature classification using Convolution Neural Network

Classifier composed of convolution layer, max pooling layer, flatten layer and fully connected layer. Each layer process the core set of feature selected by generation of feature map with high level features in pooling layer. Feature map is classified in the fully connected layer.

Conclusion

In this paper, extensive study is carried out in the classifying the Cardiomyopathy disease using geo database which contains microarray gene expression data. Gene expression data is processed with normalization technique to regularize the gene and utilizes the various feature extraction technique to extract the muted gene. Extracted gene is selected with high muted feature using feature selection technique. Selected feature is classified using machine learning and deep learning classifier to classify ischemic Cardiomyopathy ,dilated

Cardiomyopathy and neurofibromatosis Cardiomyopathy disease classes. Architecture is evaluated on experiment setup and performance measure.

References

[1] P. Luo, Y. Li, L. P. Tian, and F. X. Wu, "Enhancing the predictionof disease-gene associations with multimodal deep learning,"Bioinformatics, vol. 35, no. 19, pp. 3735–3742, 2019.

[2] D. H. Le, "Machine learning-based approaches for disease geneprediction," Briefings in Functional Genomics, vol. 19, no. 5–6,pp. 350–363, 2020.

[3] D.-H. Le and V.-T. Dang, "Ontology-based disease similaritynetwork for disease gene prediction," Vietnam Journal of ComputerScience, vol. 3, no. 3, pp. 197–205, 2016.

[4] A. Tran, C. J. Walsh, J. Batt, C. C. dos Santos, and P. Hu, "Amachine learning-based clinical tool for diagnosing myopathyusing multi-cohort microarray expression profiles," Journal ofTranslational Medicine, vol. 18, no. 1, pp. 1–9, 2020.

[5] J. Zahoor and K. Zafar, "Classification of microarray geneexpression data using an infiltration tactics optimization(Ito) algorithm," Genes (Basel), vol. 11, no. 7, pp. 1–28, 2020.

[6] R. K. Barman, A. Mukhopadhyay, U. Maulik, and S. Das,"Identification of infectious disease-associated host genesusing machine learning techniques," BMC Bioinformatics,vol. 20, no. 1, pp. 1–12, 2019.

[7] P. Popov, I. Bizin, M. Gromiha, A. Kulandaisamy, andD. Frishman, "Prediction of disease-associated mutations inthe transmembrane regions of proteins with known 3D structure,"PLoS One, vol. 14, no. 7, pp. 1–13, 2019.

[8] X. Chen, Q. Huang, Y. Wang et al., "A deep learning approachto identify association of disease-gene using information ofdisease symptoms and protein sequences," AnalyticalMethods, vol. 12, no. 15, pp. 2016–2026, 2020

[9] X. Zeng, N. Ding, A. Rodríguez-Patón, and Q. Zou, "Probability-based collaborative filtering model for predicting genediseaseassociations," BMC Medical Genomics, vol. 10, Supplement5, p. 76, 2017.

[10] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "PGCN: diseasegene prioritization by disease and gene embeddingthrough graph convolution neural networks," pp. 1–9,2019, https://www.biorxiv.org/content/10.1101/532226v1/