A Survey Paper on Advanced Approach for Classifying DNA Microarray Data using Machine Learning

Ankita Choudhary[1], Niti Shah[2]
[1]Student (Master of Engineering),[2]Assistant Professor
Computer Engineering Department, Silver Oak College of Engineering, Ahmedabad, India

Abstract- *Ensemble classification has been a frequent topic of research in recent years, especially in bioinformatics and especially true for DNA microarray data experiment and large levels of noise inherent in data. DNA Microarray data is a high-dimensional data that enables the researchers to analyze the expression of many genes in a single reaction quickly and in an efficient manner. Their characteristic such as small sample size, class imbalance, and data complexity causes it difficult to classified. The information obtained from the analysis of DNA microarray data is relevant to identify and predict illness, improve treatment and determine which genes are responsible to provoke a specific disease. This research aims at investigating and analyzing an efficient feature extraction and feature selection based approach for selecting specific features and these features use Machine learning approach for classifying DNA microarray.*

Keywords- DNA Microarray, Bioinformatics, Feature Selection, Feature Extraction, Classification

I. INTRODUCTION

Data mining is one of the newest analytical methods that have been used to serve medical science research and has been shown to be a valid, sensitive, and reliable method to discover patterns and relationships. It involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. While data mining represents a significant advance in the type of analytical tools currently available, medical research studies have benefitted from its application in many areas of interest.[10]An important problem in deoxyribonucleic acid (DNA) microarray experiments is the classification of biological samples using gene expression data. To date, this problem has received the most attention in the context of cancer research; we thus begin this work with a review of disease classification using microarray gene expression data. A reliable and precise classification of disease is essential for successful diagnosis and treatment of it. Current methods for classifying human malignancies rely on a variety of clinical, morphological, and molecular variables.[10]

II. MICROARRAY

A microarray is a multiplex lab-on-a-chip. It is a 2D array on a solid substrate (usually a glass slide or silicon thin-film cell) that tests a large amounts of biological material using high-throughput screening multiplexed, parallel processing and detection methods."Microarray" has become a general term; there are many types of microarray[12] –

- DNA microarrays
- Protein microarrays
- Antibody microarray
- Chemical compound microarray

Microarray Steps-

1. Experiment and Data Acquisition
2. Sample preparation and labeling
3. Hybridization
4. Washing
5. Image acquisition
6. Data normalization
7. Data analysis
8. Biological interpretation
9. DNA Microarray

A DNA microarrayis a collection of microscopic DNA spots attached to a solid surface.DNA microarray alsoknown as biochip or DNA chip.DNA microarray classification is a technique widely applied to discover valuable information about diseases. i.e. Cancer. It allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression). Each DNA spot contains picomoles (10^{-12} $moles$) of a specific DNA sequence, known as probes. [12]

III. DNA MICROARRAY

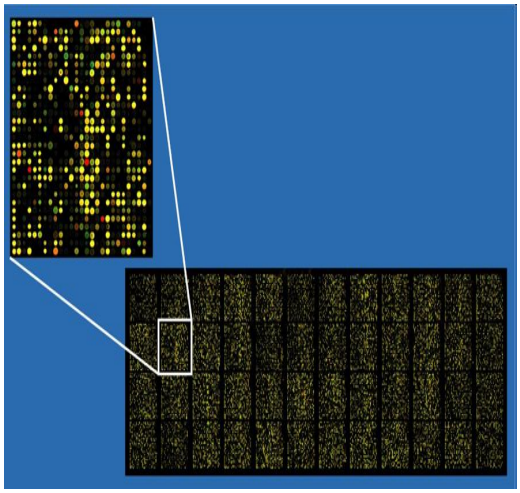

Figure 2.1 Biological Samples in 2D Arrays on Membrane[12]

GREEN represents Control DNA, where either DNA or cDNA derived from normal tissue is hybridized to the target DNA.

RED represents Sample DNA, where either DNA or cDNA is derived from diseased tissue hybridized to the target DNA.

YELLOW represents a combination of Control and Sample DNA, where both hybridized equally to the target DNA.

BLACK represents areas where neither the Control nor Sample DNA hybridized to the target DNA.

IV. RELATED WORK

4.1 Designing Artificial Neural Networks using Differential Evolution for Classifying DNA Microarrays

Author: Beatriz A. Garro, Katya Rodriguez, Roberto A. Vazquez Year: 2017 IEEE

In [1] this paper the proposed approach uses a feature selection technique based on the Artificial Bee Colony algorithm with an ANN. To test the accuracy of the proposed methodology, the author uses the Leukemia AML-ALL dataset. For performing a DNA classification task, they proposed a methodology which is divided in two main steps: the first one is focused on the selection of the genes that best describe the DNA microarray, the second one focused on designing the ANN for improving the accuracy of the classification task.

4.2 NBF: An FCA-based Algorithm to Identify Negative Correlation Biclusters of DNA Microarrays Data.

Author: AminaHouari, WassimAyadi, Sadok Ben Yahia Year: 2018 IEEE

In [2] this paper the proposed approach uses a new algorithm, called the Negative Bicluster Finder (NBF). The sighting features of the NBF stands in its ability to discover the biclusters of negative correlations using the theoretical results provided by the Formal Concept Analysis. Their results prove the NBF's ability to statistically and biologically identify significant biclusters. They present their method, called the NBF, to find biclusters with negatively-correlated patterns in microarray data. The NBF method operates in five main steps: (1) preprocessing the gene expression data matrix, (2) extracting formal concepts, (3) filtering out the obtained formal concept set's, (4) extracting negatively-correlated genes, and (5) extracting maximal negatively-correlated genes.

4.3 Mining Raw Gene Expression Microarray Data for Analyzing Synchronous and Metasynchronous Liver Metastatic Lesions from Colorectal Cancer.

Author: Hongfei, XiadongLv Year: 2016 IEEE

In [3] this paper, they present steps towards mining raw microarray data. As a case study of their approach, a public data set related to synchronous and metachronous liver metastatic lesions from colorectal cancer is then used, starting from scratch. In this work, they provide a systematic approach for mining the raw, probe-level gene expression microarray data. Major steps are summarized as follows-(1) Low-level preprocessing, (2) Additional preprocessing, (3) Quality assessment and filtering, (4) Hypothesis test, (5) Taxonomic clustering. The bottom-up analysis leads to a more rational biological discovery without employing any sophisticated algorithms. The motivation for mining raw microarray data is mainly to obtain a general comprehension of any possibly intrinsic information, to establish structure to unstructured data, and to be able to conduct serious research starting from scratch.

4.4 DNA Probe Signal Processing for Identification of Abnormal Gene Regulation and Pathogenetic Understanding – A Data Mining Approach.

Author: Hongfei Wang, WenjieCai Year: 2016 IEEE

In [4] this paper, the author present steps towards processing probe-level signal from the microarray. This paper presents a data-mining based approach to process the DNA probe signals from gene microarray data. This approach is validated through cases study of synchronous and

metachronous lever metastatic lesions from colorectal cancer, and nasopharyngeal carcinoma with cancer-free controls.

A Review of Ensemble Classification for DNA Microarray Data.

Author:Taghi M. Khoshgoftaar, David J. Dittman, Randall Wald, WaelAwada Year: 2013 IEEE

This [5] work is a review of the current state of research regarding the applications of ensemble classification for DNA microarrays. The author discusses what research thus far has demonstrated, as well as identifies the areas where more research is required. Ensemble classification is a type of technique which works well with the inherently noisy DNA microarray datasets. Benefits of ensemble classification include reduced over fitting, improved classification performance, and a reduction of bias. As of result, there has been much work regarding the applications of ensemble learning for DNA microarray datasets. Ensemble classification can be split into two categories: single classifier ensembles and multiple classifier ensembles. The majority of the research into ensemble classification utilizes single classifier approach. Two biggest problems associated with DNA microarray datasets are high dimensionality (large number of features per sample) and class imbalance (an uneven distribution of samples among the classes).

4.6LPG-PCA Algorithm and Selective Thresholding Based Automated Method: ALL and AML Blast Cells Detection and Counting.

Author:RupsaBhattacharjee, Dr. MonishaChakraborty Year: 2012 IEEE

In [6] this paper, an automated method is designed to detect ALL and AML blast cells from human microscopic blood cell images. The developed method comprises of four basic modules.

The de-noising module performs two staged noise reduction using Principal Component Analysis (PCA) and Local Pixel Grouping (LPG). The contrast enhancement section includes colour space conversion and morphological filtering based on pixel intensities. In threshold selection module, threshold value is determined using two methods, namely, Edge sensitive variational Thresholding and Ostu's Thresholding. Finally a performance evaluation is carried out in terms of accuracy based on a comparison of number of blast cells detected by manual count and those detected by this automated method.

4.7 Mining Time-Delayed Gene Regulation Patterns from Gene Expression Data.

Author: Huang- Cheng Kuo, Pei-Cheng Tsai Year: 2012 GSTF

In [7] this paper, the author proposed a modified association rule mining technique for efficiently discovering time-delayed regulation relationships among genes. By analyzing gene expression data, they can discover gene relations. Thus, the author use modified association rule to mine gene regulation patterns. Their proposed method, BC3, is designed to mine time-delayed gene regulation patterns with length 3 from time series gene expression data. They use yeast gene expression data to evaluate our method and analyze the results to show their work is efficient.

4.8 Expectation Maximization of Frequent Patterns, a Specific, Local, Pattern-Based Biclustering Algorithm for Biological Datasets.

Author: Erin J. Moore, ThirmachosBourlai Year: 2015 IEEE

In [8] this paper, the author propose an algorithm that can analyze datasets with a large attribute set at different densities, and can operate on a laptop, which makes it accessible to practitioners. Their binary Biclustering algorithm is a hybrid, axis-parallel, pattern-based algorithm that finds multiple, non-overlapping, near-constant, deterministic, binary sub matrices, with a variable confidence threshold, and the novel use of local density comparisons versus the standard global threshold. They also introduce a framework to ease comparison with other algorithms, and compare to both binary and general Biclustering algorithms using two real, and 80 synthetic databases.

4.9 Feature Selection Software Development using Artificial Bee Colony on DNA Microarray Data.

Author:WildanAndaru, IwanSyarif Year: 2017 IEEE

This [9] research aims at investigating, implementing, and analyzing a feature selection method using the Artificial Bee Colony (ABC) approach. The result is compared with other evolution algorithms, which is Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). The result is that feature selection using ABC has a better result at classification using k – Nearest Neighbor (k-NN) and Decision Tree (DT), but has a slightly higher fracture of features compared to GA and PSO algorithms.

Table 4.1 Papers summary Table

Title	Year	Approach	Open issue	My observation
Designing Artificial Neural Networks using Differential Evolution for Classifying DNA Microarrays. [1]	2017 IEEE	○ Artificial Neural Network(ANN) ○ Bioinspired Algorithms- i) Artificial Bee Colony(ABC) ii) Differential Evolution(DE) ○ Distance Classifier ○ Classification Technique - SVM	○ Performance issue in classifier ○ Classification Accuracy	To improve classification accuracy, use of different classifier such as ANN, including spiking neural network are used for testing.
NBF: An FCA-based Algorithm to Identify Negative Correlation Biclusters of DNA Microarrays Data. [2]	2018 IEEE	○ Biclustering Technique ○ Data Mining Process ○ Negative Bicluster Finder	○ Stability Measure in order to remove non – coherent concept.	Extraction of bicluster by introducing biological knowledge during extraction process.
Mining Raw Gene Expression Microarray Data for Analyzing Synchronous and Metasynchronous Liver Metastatic Lesions from Colorectal Cancer.[3]	2016 IEEE	○ Gene Expression Microarray ○ Data Mining process to dissect the genetic basis of complex diseases	○ Characterization of gene expression levels(underlying in heterogeneous and homogeneous organisms)	For mining raw microarray data is mainly to obtain a general comprehensive of any intrinsic information to establish structure to unstructured data
DNA Probe Signal Processing for Identification of Abnormal Gene Regulation and Pathogenetic Understanding – A Data Mining Approach.[4]	2016 IEEE	○ Data mining method for processing DNA Probe- level signal	○ Integrity of original data ○ Data Quality	For mining raw microarray data is mainly to obtain a general comprehensive of any intrinsic information to establish structure to unstructured data which leads to pathogenetic understanding
A Review of Ensemble Classification for DNA Microarray Data. [5]	2013 IEEE	○ Ensemble Approaches- i) Bagging ii) Boosting ○ Novel Approaches to ensemble classification- i) Single classifier ensembles ii) Multiple classifier ensemble	○ Class Imbalance (No. of instances in each class are not equal) ○ Ensemble Feature Selection ○ High Dimensionality (large no. of features per sample)	Multiple ensemble based approaches at the same time might help to improve performance, especially for difficult datasets.
LPG-PCA Algorithm and Selective Thresholding Based Automated Method: ALL and AML Blast Cells Detection and Counting. [6]	2012 IEEE	○ Connected Component Analysis ○ Local pixel grouping ○ Threshold selection module- i) Edge Sensitive Variational Thresholding ii) Otsu's Thresholding	○ Accuracy ○ Quality of Images ○ Efficiency	Image de-noising and enhancement part may be extended to generate even better performance.
Mining Time-Delayed Gene Regulation Patterns from Gene Expression Data. [7]	2012 GSTF	○ Association Rule ○ Data Mining Process ○ Gene Regulation Network	○ Multiple time unit delay ○ Cannot handle temporary ordered transactions	The time delay gene regulation patterns can be used for predicting protein- protein interaction.
Expectation Maximization of Frequent Patterns, a Specific, Local, Pattern-Based Biclustering Algorithm for Biological Datasets [8]	2015 IEEE	○ Biclustering ○ Pattern Recognization ○ Machine Learning ○ Data Mining-Association Rule ○ Unsupervised Learning ○ Computational Biology ○ Biomedical Informatics	○ Binary Biclustering algorithm are too slow ○ Non-Specific to handle biological datasets that have large number of attributes	For improving the algorithm we will let one object reference data in multiple rows of multiple tables

V. COMPARATIVE STUDY OF EXISTING METHODS

Table 5.1 Comparative Study of Existing Methods

No.	Technique Used	Strong Points	Weak Points
1	ABC Algorithm for Feature Selection [15]	Simplicity, flexibility and robustness ease of implementation High Flexibility, which allows adjustments broad applicability, complex functions	Requires new fitness tests on new algorithm parameters Higher number of objective function evaluation
2	Differential Evolution Algorithm For Feature Extraction [16]	Ability to find the true global minimum regardless of the initial parameter values. Fast and simple with regard to application and modification. Capable of providing multiple solutions in a single run. Effective on integer, discrete and mixed parameter optimization.	The convergence is unstable λ

3	ANN Classification Algorithm [17]	Storing information on the entire network Ability to work with incomplete knowledge Having Fault Tolerance Parallel processing capability	Hardware dependence Unexplained behavior of the network
4	SVM Classification Algorithm[18]	Less susceptible for over fitting of the feature input from the input items. Classification accuracy with SVM is quite impressive or high.	Multiclass items are not perfectly classified as number of items reduce gap of hyperplane.

VI. PROBLEM STATEMENT AND DEFINITION OF WORK

According to the analysis of DNA microarray, certain limitations like classifying the type of data or to identify the type of diseases are found. To overcome these reasons, it is necessary to develop a methodology that combines a robust feature selection technique with a classification algorithm for classifying DNA microarrays. Thus, the proposed model's main focus is to improve the accuracy of the classification using the Artificial Neural Network including Spiking Neural Network.

6.1Proposed System

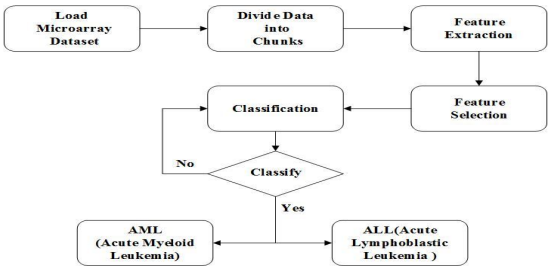

Fig 5.1: Flow Diagram

6.2 Proposed Methodology

Step 1: In the first module, the microarray gene expression data will be taken as input.

Step 2: Divide the data into chunks using segmentation technique (i.e. Apriori Algorithm).

Step 3: Feature Extraction (Features were extracted from the dataset using the following features- entropy, Standard Deviation, Skewness, Nearest Neighbour, variance or Interpolation).

Step 4: Feature Selection using Artificial Bee Colony Algorithm.

Step 5: After selecting the features, classification will be performed by using ANN with Spiking Neural Network.

Step 6: Artificial Neural Networks (ANN) are statistical learning algorithms which is one of the most effective ways for performing pattern recognition and data classification, They consist of interconnected neurons where each unit takes an input, applies a function to it and then passes the output .

Step 7:Classify the output as AML or ALL.

VIII. CONCLUSION

According to the literature analysis, Microbiological concept works in biomedical database from decades and it plays and works on many databases. To overcome the research gap in current technology, proposed method works on hybrid classification method using machine learning and feature extraction approach. With the help of this proposed work, system will detect the disease very efficiently and accurately.

REFERENCES

[1] Beatriz A. Garro and Katya Rodriguez Ciudad Universitaria , Mexico, D.F. "Designing artificial neural networks using differential evolution for classifying DNA microarrays" 978-1-5090-4601-0/17/$31.00 © 2017 IEEE

[2] AminaHouari, WassimAyadi, Sadok Ben Yahia University of Tunis, Tunisia "NBF: An FCA- based Algorithm to Identify Negative Correlation Biclusters of DNA Microarray Data" 1550-445X/18/$31.00 ©2018 IEEE

[3] Hongfei Wang, ZiadongLv Wuhan, China "Mining Raw Gene Expression Microarray Data for Analyzing Synchronous and Metachronous Liver Metastatic Lesions from Colorectal Cancer" 978-1-5090-3710-0/16/$31.00 ©2016 IEEE

[4] Hongfei Wang, WenjieCai Wuhan, China "DNA Probe Signal Processing for Identification of Abnormal Gene Regulation and pathogenetic " 978-1-5090-1345-7/16 $31.00 © 2016 IEEE

[5] Taghi M. Khoshgoftaar, David J. Dittman, Randall Wald, WaelAwada Florida Atlantic university "A Review of

Ensemble Classification for DNA Microarrays Data" 1082-3409/13 $31.00 © 2013IEEE

[6] RupsaBhattacharjee , Dr. MonishaChakraborty West Bengal , India "LPG-PCA Algorithm and selective Thresholding based Automated Method: ALL & AML Blast Cells Detection and Counting" 978-1-4673-4700-6/12/$31.00 © January 2012 IEEE

[7] Huang-Cheng Kuo and Pei- Cheng Tsai "Mining Time-delayed Gene Regulation Patterns from Gene Expression Data" ©2012 GSTF

[8] Erin J.Moore, ThirmachosBouriaiMember , IEEE "Expectation Maximization of Frequent Patterns, a Specific, Local, Pattern-Based Biclustering Algorithm for Biological Datasets" 1545-5963 ©2015 IEEE

[9] WildanAndaru, IwanSyarif, Ali RidhoBarakbahSubrabaya, Indonesia "Feature Selection Software Development Using Artificial Bee Colony On DNA Microarray Data" 978-1-5386-0716-9/17/$31.00 ©2017 IEEE

Website-

[10] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," San Fr. CA, itd Morgan Kaufmann, p. 745, 2012.

[11] http://dataminingwarehousing.blogspot.com/2008/10/data-mining-steps-of-data-Mining .html

[12] https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[13] https://www.worldscientific.com/doi/pdf/10.1142/S0219622006002 258

[14] https://www/informatics.indiana.edu/rocha/publications/ps/ mrochaDataMiningMAs.pdf

[15] https://www.slideshare.net/VinayakNayak5/abc-algorithm-49203658

[16] http://shodhganga.inflibnet.ac.in/bitstream/10603/10244/7/07_chapter%202.pdf

[17] https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel

[18] http://www.cs.uky.edu/~jzhang/CS689/PPDM-Chapter2.pdf