International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 4 Apr. 2021, pp: 237-242 www.ijaem.net ISSN: 2395-5252

Microarray Baseddisease Prediction Using Deep Learning Techniques

V Gokulakrishnan[1], K Madhubala[2], R Selvasarathi[3], R Dhivya[4]

[1]*Assistant Professor, Department of CSE, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamil Nadu,India*

[2, 3, 4]*UG Students, Department of CSE, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamil Nadu, India*

Submitted: 15-04-2021 Revised: 28-04-2021 Accepted: 30-04-2021

ABSTRACT: The DNA microarray technology has modernized the approach of biology research in such a way that scientists can now measure the expression levels of thousands of genes simultaneously in a single experiment. Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. Diseases classification with gene expression data is known to include the keys for addressing the fundamental harms relating to diagnosis and discovery. The recent introduction of DNA microarray technique has complete simultaneous monitoring large number of gene expressions possible. With this large quantity of gene expression data, experts have started to discover the possibilities of disease classification using gene expression data. Quite a large number of methods have been planned in recent years with hopeful results. In order to gain insight into the disease classification difficulty, it is necessary to get a closer look at the problem, the proposed solutions and the associated issues all together. This present a comprehensive searching method, clustering method and classification method such as Pattern similarity search, Spatial Expectation Maximization, nearest neighbour classification and estimate them based on their evaluation time, classification accuracy and ability to reveal biologically meaningful gene information. Based on our multiclass classification method to diagnosis the diseases such as Cancer (Lung, Blood, Breast and Skin) diseases and other diseases and also find severity levels of diseases and also prescribe the medicine for affected diseases. This experimental results show that classifier performance through graphs with improved accuracy.

Keywords: Gene selection, cancer microarray data, cuckoo search, multi-objective, evolutionary operators.

I. INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing application software's are inadequate to deal with them. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics or certain other advanced data analytics methods that extract value from data. Scientists, business executives, practitioners of medicine, advertising and governments like regularly meet difficulties with large data-sets in areas including Internet search, finance, urban informatics and business informatics.

II. RELEVANT WORK

Big data can be described by the following characteristics: Volume, Variety, Velocity, Variability, and Veracity.

(i)Volume – The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

(ii)Variety – The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

(iii)Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data. Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

(iv)Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

(v) Veracity - The quality of captured data can vary greatly, affecting accurate analysis. Factory work and Cyber-physical systems may have a 6C system.

Application of Big Data:

1) **Government:** The use and adoption of big data within governmental processes allows efficiencies in terms of cost, productivity and innovation, but does not come without its flaws. Data analysis often requires multiple parts of government (central and local) to work in collaboration and create new and innovative processes to deliver the desired outcome.
2) **International Development:** Research on the effective usage of information and communication technologies for development (also known as "ICT4D") suggests that big data technology can make important contributions but also present unique challenges to international development. Advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security and natural disaster and resource management.

- A major practical application of big data for development has been "fighting poverty with data". In 2015,Blumenstock and colleagues estimated predicted poverty and wealth from mobile phone metadata and in 2016 Jean and colleagues combined satellite imagery and machine learning to predict poverty.
- Thematic coverage: including areas that were previously difficult or impossible to measure.
- Geographical coverage: our international sources provided sizable and comparable data for almost all countries, including many small countries that usually are not included in international inventories.
- Level of detail: providing fine-grained data with many interrelated variables and new aspects, like network connections.
- Timeliness and time series: graphs can be produced within days of being collected.

3) **Education:** A McKinsey Global Institute study found a shortage of 1.5 million highly trained data professionals and managers and a number of universities including University of Tennessee and UC Berkeley, have created masters programs to meet this demand.
4) **Media:** To understand how the media uses big data, it is first necessary to provide some context into the mechanism used for media process. It has been suggested by Nick Couldry and Joseph Turow that practitioners in media and advertising approach big data as many actionable points of information about millions of individuals.
5) **Insurance:** Health insurance providers are collecting data on social "determinants of health" such as food and TV consumption, marital status, clothing size and purchasing habits, from which they make predictions on health costs, in order to spot health issues in their clients.
6) **Information Technology:** Especially since 2015, big data has come to prominence within business operations as a tool to help employees work more efficiently and streamline the collection and distribution of information technology (IT). The use of big data to resolve IT and data collection issues within an enterprise is called IT operations analytics (ITOA).

III. GENE BASED ANALYTICS

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene. The spots are printed on to the glass slide by a robot or are synthesized by the process of photolithography. Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes and subtypes of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available. Furthermore, co-expressed genes in the same clusterare likely to be involved in the same cellular processes and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cis-regulatory elements to be proposed. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches.

CHALLENGES IN GENE CLUSTERING:

Due to the special characteristics of gene expression data and the particular requirements from the biological domain, gene-based clustering presents several new challenges and is still an open problem. First, cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis. For example, a clustering algorithm which can accurately estimate the “true” number of clusters in the data set would be more favored than one requiring the pre-determined number of clusters. Second, due to the complex procedures of microarray experiments, gene expression data often contain a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise. Third, our empirical study has demonstrated that gene expression data are often “highly connected” and clusters may be highly intersected with each other or even embedded one in another. Therefore, algorithms for gene-based clustering should be able to effectively handle this situation. Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters (e.g., which clusters are more close to each other, and which clusters are remote from each other) and the relationship between the genes within the same cluster (e.g., which gene can be considered as the representative of the cluster and which genes are at the boundary area of the cluster). A clustering algorithm, which can not only partition the data set but also provide some graphical representation of the cluster structure, would be more favored by the biologists.

STEPS:

Datasets Acquisition

A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index and make the data available to other applications for analysis and interpretation. Then, upload the datasets. The dataset may be microarray dataset. A microarray database is a repository containing microarray gene expression data. Then implement preprocessing steps to eliminate the irrelevant symbols.

Median Estimation

The median is the value separating the higher half from the lower half of a data sample (a population or a probability distribution). For a data set, it may be thought of as the "middle" value. The median is a commonly used measure of the properties of a data set in statistics and probability theory. The basic advantage of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed so much by a small proportion of extremely large or small values, and so it may give a better idea of a "typical" value. Then split the gene symbols into 2^2 combinations. And eliminate the header symbol for future calculation. Declare the predefined Template as CTAG and calculate frequency count for each symbol.

PSO Algorithm

In PSO algorithm, can analyze coverage of the data before clustering begins. And propose an algorithm, which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering. It has two distinct phases: one of transferring an object from one cluster to another and the other of amalgamating the single member cluster with it’s the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested.

Disease Prediction

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings.Then implement K nearest neighbour algorithm to classify the various types of diseases from gene expression.Classification is done with the help of KNN classifier. In the recent years, KNN classifiers have established excellent performance in a variety of pattern recognition troubles. The input space is planned into a high dimensional feature space. Then, the hyper plane that exploits the margin of separation between classes is constructed. The points that lie closest to the decision surface are called support vectors directly involves its location. When the classes are non-separable, the optimal hyper plane is the one that minimizes the probability of classification error. Initially input image is formulated in feature vectors. Then these feature vectors mapped with the help of kernel function in the feature space. And finally division is computed in the feature space to separate out the classes for training data.The KNNs algorithm separates the classes of input patterns with the maximal margin hyper plane. This hyper plane is constructed as:

$$f(x) = \langle w, x \rangle + b$$

Where x is the feature vector, w is the vector that is perpendicular to the hyper plane and $b\|w\|^{-1}$ specifies the offset from the beginning of the coordinate system. To benefit from non-linear decision boundaries the separation is performed in a feature space F, which is introduced by a nonlinear mapping φ the input patterns.This mapping is defined as follows:

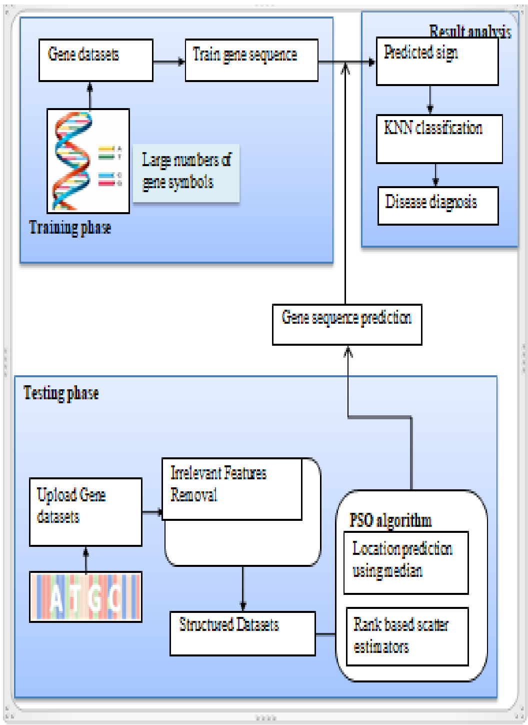
$$\langle \varphi\ (x_1), \varphi(x_2) \rangle = K(x_1, x_2)\ \forall (x_1, x_2) \in X$$

for some kernel function K (·, ·). The kernel function represents the non-linear transformation of the original feature space into the F.

Severity Analysis

Using multi class classification algorithm to classify the severity level of diseases using classified data count. If count is more than threshold means, provide severity as high and count is less than threshold means, consider as normal. Then provide prescription to patients according to the diseases.

IV. SYSTEMDESIGN



V. SOFTWARE TESTING

Software testing is a method of assessing the functionality of a software program. There are manydifferent types of software testing but the two main categories are dynamic testing and static testing. Dynamic testing is an assessment that is conducted while the program is executed; Static testing, on the other hand, is an examination of the program's code and associated documentation. Dynamic and static methods are often used together.Testing is a set activity that can be planned and conducted systematically. Testing begins at the module level and work towards the integration of entire computers based system.

Testing Objectives

There are several rules that can serve as testing objectives, they are

- Testing is a process of executing a program with the intent of finding an error.
- A good test case is one that has high probability of finding an undiscovered error.
- A successful test is one that uncovers an undiscovered error.

If testing is conducted successfully according to the objectives as stated above, it would uncover errors in the software. Also testing demonstrates that software functions appear to the working according to the specification, that performance requirements appear to have been met.Tests for correctness are supposed to verify that a program does exactly what it was designed to do.

Tests used for implementation efficiency attempt to find ways to make a correct program faster or use less storage. It is a code-refining process, which reexamines the implementation phase of algorithm development. Tests for computational complexity amount to an experimental analysis of the complexity of an algorithm or an experimental comparison of two or more algorithms, which solve the same problem.

The data is entered in all forms separately and whenever an error occurred, it is corrected immediately. A quality team deputed by the management verified all the necessary documents and tested the Software while entering the data at all levels. The development process involves various types of testing. Each test type addresses a specific testing requirement. The most common types of testing involved in the development process are:

- Unit Test.
- Functional Test
- Integration Test

Unit Testing

The first test in the development process is the unit test. The source code is normally divided into modules, which in turn are divided into smaller units called units. These units have specific behavior. The test done on these units of code is called unit test. Unit test depends upon the language on which the project is developed. Unit tests ensure that each unique path of the project performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Functional Testing

Functional test can be defined as testing two or more modules together with the intent of finding defects, demonstrating that defects are not present, verifying that the module performs its intended functions as stated in the specification and establishing confidence that a program does what it is supposed to do.

Integration Testing

In integration testing modules are combined and tested as a group. Integration Testing follows unit testing and precedes system testing. Testing after the product is code complete. Betas are often widely distributed or even distributed to the public at large in hopes that they will buy the final product when it is released.

VI. CONCLUSION

Microarray is an important tool for cancer classification at the molecular level. It monitors the expression levels of large number of genes in parallel. With large amount of expression data obtained through microarray experiments, suitable statistical and machine learning methods are needed to search for genes that are relevant to the identification of different types of disease tissues. In this thesis,have proposed a hybrid gene selection method, which combines a PSO methods and KNN classification to achieve high classification performance. The method was designed to address the importance of gene ranking and selection prior to classification, which improves the prediction strength of the classifier. The project focused on promising accuracy results with very few number of gene subsets enabling the doctors to predict the type of cancer. The results on various disease datasets shows the importance of the same classifier used for both the gene selection and classification can improve the strength. Then provide severity level for each classified diseases.

Future work includes partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are tightly coupled with strong association to the sample categories. We can extend the work to implement various classification algorithms to improve the accuracy rate at the time of disease prediction.

REFERENCES

[1] Sun, Lin, et al. "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification." Information Sciences 502(2019) : 18-41.

[2] Tandel, Gopal S, et al. "A review on a deep learning perspective in brain cancer classification." Cancers 11.1(2019): 111.

[3] Huang Shujun, Nianguang Cai. “Applications of support vector machine (SVM) learning in Cancer Genomics.” Cancer Genomeics & Proteomics 15(2018): 41-51.

[4] Danaee Padideh. “A deep learning approach for cancer detection and relevant Gene identification.” Biocomputing (2017):219-229.

[5] Darshan S.Chandrasekar. “A portal for facilitating Tumor Subgroup Gene Expression and Survival analyses.” UALCAN 19.8(2017):649-658.

[6] Li, Yuanyuan, et al. "A comprehensive genomic pan-cancer classification using The

Cancer Genome Atlas gene expression data." BMC genomics 18.1(2017): 1-13.

[7] Lu, Huijuan, Junying Chen. “A hybrid feature selection algorithm for gene expression data classification.” Neurocomputing 10.15(2017):1-7.

[8] Radovic, Milos, et al.“Minimum redundancy maximum relevance feature selection approach for temporal gene expression.” Bioinformatics 18.9(2017):1-14.

[9] Shiozawa, Yusuke, et al."Gene expression and risk of leukemic transformation in myelodysplasia." Blood 130.24(2017): 2642-2653.

[10] Tarek, Sara, Reda Abd Elwahab, and Mahmoud Shoman. "Gene expression based cancer classification." Informatics 18.3(2017): 151-159.

AUTHORS

First Author –V Gokulakrishnan M.E.,M.B.A.,Assistant Professor, Department of CSE, Dhanalakshmi Srinivasan Engineering college, Perambalur, Tamil Nadu, India.

Second Author –K Madhubala, Department of CSE, Dhanalakshmi Srinivasan Engineering College,Perambalur, Tamil Nadu, India.

Third Author –R Selvasarathi,Department of CSE, Dhanalakshmi Srinivasan Engineering College,Perambalur, Tamil Nadu, India.

Fourth Author–R Dhivya,Department of CSE, Dhanalakshmi Srinivasan Engineering College,Perambalur, Tamil Nadu, India.