Classification of Health Disorder Based On DNA Technology

Sudha V [1], Dr.Girijamma[2], Pragati S [3]

Assistant professor, Department of Information and Science, RNSIT, sudharam07@gmail.com
Professor, Department of Computer Science and Engineering, RNSIT, girijakasal@gmail.com
M.Tech (CNE) Student, Department of Information and Science, RNSIT, pragu.vijaya@gmail.com

---***---

Abstract - *Computational biological researchers applied various paradigms like bioinformatics paradigms in microarray technology on gene expression data. A gene matrix has been obtained as a form of expression data in microarray with rows as large count of gene and time points / disease states representing the columns. Bioinformatics job have been made critical with these data sets. To distinguish the features within disparate variety of samples like natural or the affected, a section of genetic code is selected from an enormous riotous datasheet. And the severity of the disease is classified based on the gene profile samples gathered from a placenta of humans. In the process of detecting and preventing the cancer recognition of genetic and environmental factor plays a crucial role. Accordingly cancer risk perdition system is built with clustering, expected maximization algorithm with k-nearest neighbor method. In identification of cancer classification clustering and perdition technology are used. A number of gene selection data are grouped which is termed as critical for cancer. The normal and diseased data set have been selected , the gathered data is preprocessed, fed into the data base and classified to yield significance patterns using expected maximization algorithm the data is clustered to separate cancer and non-cancerous patients.*

Key Words: ***Microarray Technology, Gene Expression Data, clustering, K-Nearest Neighbour, Expected Maximization Algorithm***

1. INTRODUCTION

The studies on gene have made a way in understanding the transcriptional dynamic on a cell under many biological stresses. The microarray has a robust and amenable system of recording transcription profile. Therefore diagnosing from gene data gives more solid and proficient and repeatable diagnosis than authentic histopathology method. Gene is a path in DNA that encodes activity. A chromosome contains long strand of DNA which in turn has many gene. It can be differentiated using genetic codes. A genetic code is a sequence of amino acids. There are three sequential basis of DNA code in chromosomes. The physical and behavioral attribute of a body is decided by gene present in chromosomes. The protein production in gene may vary depending on pressure, temperature and different metabolic concentration in the cellular environment. This circumstance is known as gene expression. Transcription and translation are the main metabolic reactions occur in gene in which transcription produces the RNA. The entire mRNA is isolated in the gene sample to classify between the normal and diseased gene. Transcriptase enzyme is used to produce complementary DNA form isolated mRNA. Steady states and time series are different types of information from gene profile which is expressed as a form of matrix. The data in periodic non-parallel gene expression in column represents the time points. In steady state the tissues under observation are the samples of steady state data. The expression levels are mapped as floating point number before it is being used in bio-informatics task in a certain range. Through biological developments a new era of gene expression data are produced which results to increase in number of rows and column in data matrix. A gene is simpler functional and physical unit of heredity. Gene acts as an instruction to make molecules called proteins. The size varies from hundred to more than 2 million DNA bases. The human DNA has 20,000 to 25,000 genes.

The gene contains the copy of each gene from your parents. Some genes are common in all people and less than 1 percent of the total number people are slightly different. The gene with minor difference is called Alleles. This minor difference contributes to uniqueness of the person. The genes in the DNA are responsible for different function. Group of gene in a DNA make up chromosomes. It is present inside the nucleus of the cell. Chromosomes have uniqueness in the structure. For a living creature the genetic information determine the overall features of it. A particular character is influenced by large quantity of gene. From the polygenic character to the disease that affects a person is decided by the gene. Even a simpler unit of ancestry is caused by a gene in the living-organism. Genes are derived from our ancestors. A gene will incorporate all the data required for building and conserving a new cell and also these genetic details are given offspring. A pair of chromosomes present in a particular gene is extracted from their ancestor. The genome details play an important role in counter measures and aid of any disease. Matrons with tumor suppressor have larger probability of getting breast cancer. To know whether the genome is affected by a disease the preliminary analysis are taken. Gene selection is an important process in disease classification. In the process a small subset of gene is sufficient for an efficient outcome. The different classes of samples are differentiated using K-nearest neighbor method. The selection of subset of gene is based on the prediction from the large noisy data set. There are many methods in gene expression profile but the number of gene involved exceeds several thousands in the process of selection. Relatively a small size of data set will be available for the classification.

Accordingly the classification quality will be decreased because of redundant gene. Hence to reduce the noisy gene, the pre-filtering approach is used to overcome the complication in selection. Here the gene selection process is based on an unsupervised technique. One of the unsupervised method is clustering where a set of patterns are usually grouped into clusters. The patterns within the clusters are same but it is different between the clusters. The similarities between two patterns X and Z are identified by the Euclidean distance D where D=||X-Z||. Bigger the distance between X and Z lesser is the similarities between them and vice versa. One of pattern recognition application used here is KNN classifier. And expected maximization algorithm is used to differentiate between the normal and diseased gene. Gene gives complete structure of body and cellular environment which paved a way in identification of different diseases. One of such deadly disease is cancer. It is a disease in which unusual cells divides unmanageably and destroy the body tissue. Due to the changes that occur in a gene over a lifetime of a person may cause cancer. It is not a genetic disease but in rare cases it is so. To identify this sort of disease in the initial stages there are many techniques, one such is expected maximization which will be explained in the upcoming chapters.

2. Proposed System

In proposed system the gathered dataset is used for classification of a disease called cancer. Here the prediction system is enhanced with expected maximization algorithm, KNN and clustering methods to find the best outcomes.

2.1 Advantages of proposed system

- Large datasets with respect chromosomes representations which should reflect the bioinformatics on gene expression data.
- Noise predictions.
- Data points.
- All unique items discovery and assign the weight age for all the items.
- Express the invert the matrix with all the data.
- Put sugar levels for maximum frequency generated item sets.

2.2 System Architecture

The process of characterizing the data's, architectus, schedules, peripherals and interfaces for application software is known as system design. The application of ideological theory is the development. The areas of systematic analysation, architectus and handlings have overlapping. In the process of determination and system development for convincing the user's prerequisite is termed as system design. The Fig-1 shown below represents the system architecture diagram for the classification of health disorders.

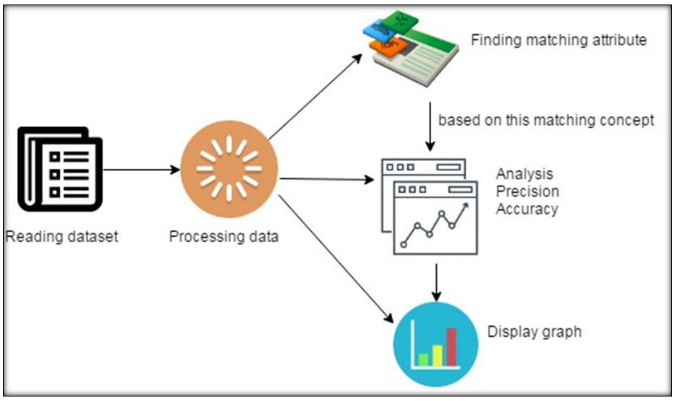


Fig-1: System Architecture

The system architecture includes reading dataset, processing data, finding matching attribute and prediction. The datasets required for the processing will be stored in .xls file format. Once the data is obtained then using the k-value the properties of the selected data is extracted for the processing and based on which the suitable matching attributes are recognized and the values are predicted.

2.3 Sequence-flow Diagram

The Fig-2 describes the sequence flow diagram of the execution. Here the interaction is between the user, dataset, matching count and analysis. The read command is performed between user and dataset, count values and detail interaction is between the dataset along with matching count. Finally the calculated accuracy, precision and graph of matching count is among the user and analysis module.

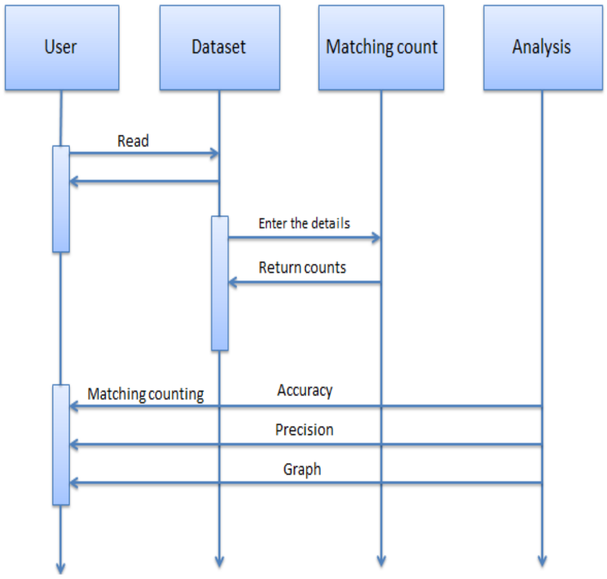


Fig -2: Sequence Diagram

3. Requirement Specification

3.1 Software Requirements

Software requirements deal with defining software resource requirements and prerequisites that need to be installed on a computer to provide optimal functioning of an application. These requirements or prerequisites are generally not included in the software installation package and need to be installed separately before the software is installed.

- JAVA 1.4 or higher
  - o Java Swing – front end
  - o JDBC –Database connectivity
  - o UDP-User Datagram Protocol
  - o TCP-Transmission Control Protocol
  - o Networking-Socket programming
- ORACLE –Back end
- Windows 98 or higher-Operating System

3.2. Hardware Requirements

The most common set of requirements defined by any operating system or software application is the physical computer resources, also known as hardware, A hardware requirements list is often accompanied by a hardware compatibility list, especially in case of operating systems. An HCL lists tested, compatible, and sometimes incompatible hardware devices for a particular operating system or application. The following sub-sections discuss the various aspects of hardware requirements. All computer operating systems are designed for particular computer architecture. Most software running on x86 architecture define processing power as the model and the clock speed of the CPU. Many other features of a CPU that influence its speed and power, like bus speed, cache, and MIPS are often ignored. This definition of power is often erroneous, as AMD Athlon and Intel Pentium CPUs at similar clock speed often have different throughput speeds.

- 10GB HDD(min)
- 128 MB RAM(min)
- Pentium P4 Processor 2.8Ghz(min)

4. Experimental Result

4.1 Datasets

We used microarray datasets to evaluate our method. These datasets are used commonly for sample classification and gene clustering research.

4.2 Results and Discussion

We conducted a set of experiments to demonstrate the applicability and effectiveness of the proposed framework. For this purpose, we have used different existing software tools for realizing the feature reduction models and the classification tasks. The proposed work is compared with the existing work and the graph is plotted for the same in fig-3.

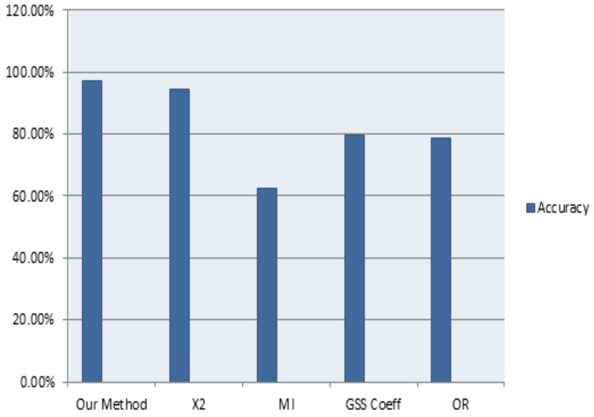
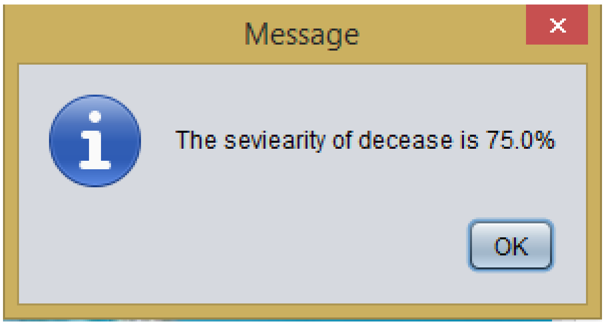


Fig-3:Output Graph



The seviearity of decease is 75.0%

Fig-4: Represents the severity of disease

The fig-4 shows that the level of severity that is affected to a particular gene data selected for the process.

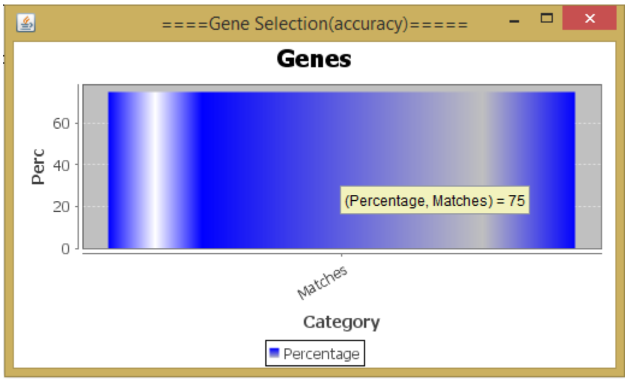


Fig-5: Percentage of accuracy

The fig-5 depicts that the accuracy of prediction percentage in classifying the dataset selected for the processing

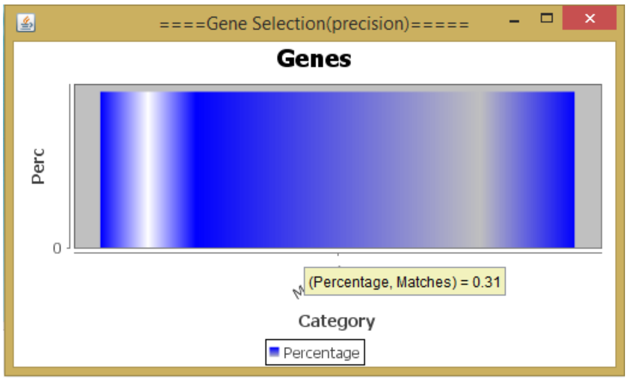


Fig-6: Percentage of precession

The figure7.13 shows that the precession of prediction percentage in classification health disorders.

5. CONCLUSIONS

Cancer has been a major cause of death worldwide. The most efficient way to prevent cancer deaths is to identify it in earlier stages. Many people do not check it in earlier stage because due to the cost involved in taking several tests for diagnosis. This identification system may provide easy and a cost effective way for identifying cancer and may play an important role in earlier diagnosis process for various types of cancer and provide effective preventive strategy. Here the selection process is based on clustering along with KNN and for perdition expected maximization is used. The outcome of the execution results the severity of the disease and finally a graph is plotted based on accuracy and precision.

This work can be stretched out to SAAS on W3C for web4.0 designs. Coming to general stubs and skeletons both sides the calculation part can be inundated taking into account augmentation of both information sets and methodologies relocation. The SOA design movement of this work is effortlessly supportive for further preparing of information like pre-handling and quick examination of different models and as described information. Tossing the information to cloud and on W3C of these sorts of datasets constantly touchy however for security and analysation reason stubs and skeletons keeps up security to prepare the information in mining and arrangement levels. For example disclosure this work can be stretched out from KNN level without further handling of choice trees. This is supportive in current work for different purposes.

REFERENCES

1) B.A. Pierce, *Genetics: A Conceptual Approach*4thed, W. H. Freeman and Company, 2012

2) B. Lewin, *Genes VIII*, 8thed, Published by Pearson Prentice Hall,2004.

3) D. Stekel, *Microarray Bioinformatics*, 1sted, Cambridge University Press, 2003.

4) R. J. Lipshutz, S. P.A. Fodor, T. R. Gingeras et al. "High density synthetic

oligonucleotide arrays", nature genetics, vol. 21, no. 1, pp. 20 – 24, 1999.

5) D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays", Nature, vol. 405, pp. 827 – 836, 2000.

6) D. L. Donoho, "High-dimensional data analysis: the curses and blessings of dimensionality," in American Mathematical Society Conf. Math Challenges of the 21st Century, 2000.

7) C. Tsai, C. Chen, T. Lee, I. Ho, U. Yang, "Gene Selection for Sample Classifications in Microarray Experiments", DNA and Cell Biology, vol. 23, no. 10, pp. 607–614, 2004.

8) H. Al-Mubaid, N. Ghaffari, "Identifying the Most Significant Genes from Gene Expression Profiles for Sample Classification", IEEE International Conference on Granular Computing, pp. 655 – 658, 2006.

9) K. E. Basford, G. J. Mclachlan, S. I. Rathnayake, "On the Classification of Microarray Gene-expression Data", Brief Bioinform. vol. 14, no. 4, pp. 402 – 410, 2013.

10) P. J. Williams, F. B. Pipkin, "The genetics of preeclampsia and other hypertensive disorders of pregnancy", Best Practice & Research Clinical Obstetrics and Gynecology, vol. 25 pp. 405– 417, 2011.

11) H.Laivuori, "Genetic aspects of preeclampsia", Front Biosci, vol. 12, pp. 2372-2382, 2007.

12) L. Li, C. R. Weinberg, T. A. Darden et al. "Gene Selection for Sample Classification based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN method", Bioinformatics (Oxford Journal), vol. 17, no. 12, pp. 1131 – 1142, 2001.

13) L Li and C. R. Weinberg, "Gene Selection and Sample Classification Using a Genetic Algorithm and k –Nearest Neighbor Method" In: A Practical Approach to Microarray Data Analysis, pp 216-229, Springer, 2003.

14) E. Tejera, J. Bernardes, I. Rebelo, "Co-expression network analysis and genetic algorithms for gene prioritization in preeclampsia", BMC Medical Genomics,ol. 6, no. 51, 10 pages, 2013.

15) S. Deegalla, H. Bostrom, "Classification of Microarrays with kNN: Comparison of Dimensionality Reduction Methods" In: *Intelligent Data Engineering and Automated Learning,* pp. 800 – 809, Springer, 2007.