

Cancer Therapeutics Based on Gene Expression Data: A Distributed Approach

Thushara s^{1*}, Chinnu Ravi²

^{1*}PG Scholar, Dept. of Computer Science and Engineering

²Asst. Professor, Dept. of computer science and engineering
TKM Institute of Technology, Kollam, India

Abstract – Several deaths are caused due to cancer. Cancer detection and classification is an important issue in the field of bioinformatics. Machine learning technique such as classification and data mining technique such as clustering are previously applied for identifying cancer using gene expression data. But the technique used is time consuming. In this paper we implement a distributed clustering algorithm by which time consumption is lesser. In this algorithm we planned to implement hierarchical clustering in a distributed manner.

Index terms – Cancer, machine learning, data mining, clustering, classification, gene expression data, distributed clustering algorithm

1 INTRODUCTION

Class of diseases that is characterized by uncontrollable cell growth is the cancer. Any body part may be affected with cancer. Causes of cancer still remains unknown but several factors are not avoided completely. Cancer needs to be cured earlier. If it reaches the extreme state then the patient will be severely affected. Earlier day's treatment of cancer was done clinically. But this kind of detection is possible only at an extreme state, so curing is not effective. We have to find out which property of gene has caused the disease so that early detection may become possible. For this purpose gene, the property of gene, DNA etc. has to be analyzed in detail.

2 DNA

DNA the deoxyribonucleic acid found in all living organism including many viruses is the molecule encoding the genetic instructions. The double helical structure of DNA is formed by two biopolymer strands. The polynucleotide is the name given to the strands which consists of several simple nucleotides. Guanine, adenine, thymine, cytosine are the nucleobase found in these nucleotides. Biological information storage is the main function of DNA. Long structure like organization of DNA is called chromosomes. James Watson and Francis Crick are the two scientists who discovered the double helical structure of DNA. Mutations in DNA will lead to cancer. Many mutagens will damage the DNA sequence. DNA damage varies depends on type of mutagens. Several types of mutagens are identified they are oxidizing agents, alkylating agents, electromagnetic radiation, ultraviolet rays and x-rays. If the double strand breaks then it is difficult to repair the strand causing insertion, deletion,

mutations, and chromosomal translocations and finally leads to cancer.

3 DIAGNOSIS USING MICROARRAYS

Diagnosis using microarray is a complicated task. The data in microarray contains an image. This image is converted into the form of a gene expression matrix. Virtual lab on a chip is called microarray. 2D array on a solid platform is the composition of microarray. Solid platform may be a glass slide; the purpose of this microarray is for biological reference study. There are many types of microarrays i.e. DNA microarrays, protein microarrays, peptide microarray, tissue microarray, cellular microarrays etc.

4 EXISTING METHOD

In the existing system cancer detection and classification was done based on gene expression data. Machine learning technique and data mining technique like clustering and classification was used for the purpose. For finding which property of gene has caused the disease association rule has been derived. Extension of clustering and classification give rise to the association rule. After classification several association rules can be formed. From this association rule best rule was taken and that best rule was stored as knowledge. When sample was given knowledge was processed then tested to get the result as yes or no. There are some disadvantages in this system. The large microarray dataset is not processed by a system. And the system was time consuming.

5 PROPOSED METHOD

The field of computer science that studies distributed system is called distributed computing. Software system components in the networked computer communicate and coordinate their action by passing messages is called a distributed system. In distributed we implement a distributed clustering algorithm. In this algorithm, we planned to implement hierarchical clustering in a distributed manner. According to this, initial dataset is divided into N dataset. And these N datasets are stored in 'N' cluster nodes. Each of these cluster nodes perform clustering on its own dataset and form a partial hierarchical cluster. And special nodes CMN (Cluster Merge Node) which combine these N partial hierarchical into a single cluster and from the several expressions are derived. From their best expression is taken. And this best expression

is stored as knowledge. Rule saved as knowledge was processed and then tested to get the result as yes or no.

A. SYSTEM ARCHITECTURE

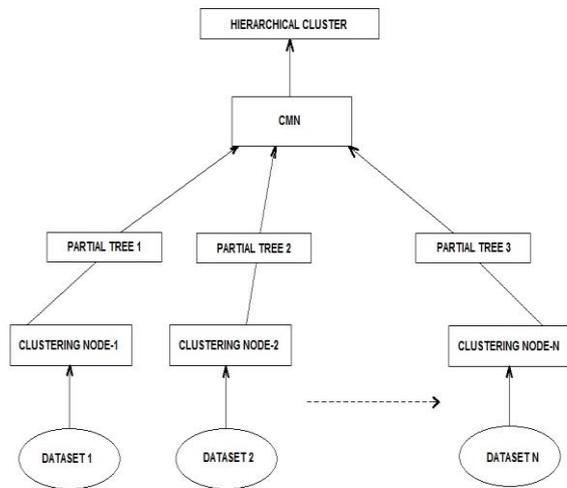


Fig1: system architecture

Here the initial data set is divided into N dataset. This N Data Set performs clustering using K-means clustering. The clustering in which grouping of objects occurs where variability is small within and large across the clusters. In k-means K number of clusters is there. The partial trees obtained after clustering are combined by cluster merge node into a single partial hierarchical cluster. Partial tree consist of diseased tree and a sub tree.

ALGORITHM 1: SERVER SIDE

Expression tree (ET)

- Set of expressions formed
- Removed tree

1. Split the dataset into N blocks.
2. Send to N blocks
3. Declare ET[n]
4. For each node i
Receive ET[i]
5. Declare expression list, EL
6. For each i=1 to N
Exp=ET[i].expression
Add each element in exp to EL, and perform
Optimization
7. for each ET
Begin
Eliminate node in ET using EL
End
8. Build a tree, T using all the remaining nodes in ET
9. Find expressions in T
10. Add EL
11. Best rule finding

ALGORITHM 2: NODE SIDE

1. Receive dataset.
2. Preprocessing.
3. Clustering.
4. Classification.
5. Find expression tree.
6. Send expression tree to the server.

The partial formed after clustering diseased tree is taken and the non-diseased sub tree is there. Expression tree is composed of set of expressions and removed tree. At the server side data set is divided into N dataset and that data set is send to N blocks. After clustering several partial trees are obtained. Nodes are evaluated and same expressions are removed. I.e. if the tree contains same expression to that of another tree that expressions are removed. And if one tree is composed of some expressions in another tree that expressions are removed. After removing duplicates best expression is taken as the rule.

6 RELATED WORKS

Kyu-Baek [1] Hwang proposed the machine learning techniques like Bayesian networks, neural trees and radial basis function for the analysis of gene expression data. Benny Y. M Fung [2] proposed a MIF algorithm for classifying multi type gene expression cancer types. Herbert Pang [3] identified genes that differentiate between oestrogen negative and oestrogen positive patients that suffer from breast cancer in Africa and America. P E Lonning [4] proposes a method for finding prognostic and predictive factors in breast cancer therapy. Jungui Chen [5] used pattern recognition and data mining technique for identifying colon cancer using gene expression data. Shital shah [6] introduced an integrated gene search algorithm for the analysis of genetic expression data. Jin hyuk hong [7] proposed an ensemble approach for genetic programming. Andrew campen [8] proposed a gene to disease expression mapper (D-GEM) a data mining tool for identifying human primary disease gene. Johan Ingvarsson proposed [9] a recombinant scFv antibody microarray in an attempt to derive sera derived from adenocarcinoma patients versus healthier samples. Pascale anderle [10] proposed a generalized work flow scheme typical for microarray experiments using two samples related to cancer research. Roberto Ruiz [11] proposed a new heuristic to select relevant gene subset in order to further use them for the classification task. Olivier Gevaert [12] identified the impact of DNA microarray and Omic technology on cancer patients.

7 CONCLUSION

The work proposed a distributed approach for cancer therapeutics based on gene expression data. Here we implement hierarchical clustering in a distributed manner. Since the large dataset is processed in different machines time delay can be avoided. For writing association rule time consumption is avoided. Since the system is working in a distributed manner overhead will occur.

8 REFERENCES

[1] ShauryaJauhari and S.A.M Rizvi, "Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest", IEEE Transactions on Computational Biology and Bioinformatics, vol .11, no.3 533-547, June 2014.
 [2] Kyu-Baek Hwang, Dong-Yeon Cho, "Applying Machine Learning Techniques to Analysis of Gene Expression Data: Cancer Diagnosis", Seoul 151-742, 2002-Springer.

- [3] Shital Shah, Andrew Kusiak, "Cancer Gene Search with Data-mining and Genetic Algorithms", *Computers in Biology and Medicine* 37,251-261, 2007.
- [4] P E Loaning, T Sorlie, "Microarrays in Primary Breast Cancer- Lessons from Chemotherapy Studies", *Endocrine Related Cancer* 8, 259-263, 2001.
- [5] Johan Ingvar son, ChristerWingren, "Detection of Pancreatic Cancer Using Antibody Microarray-based Serum Protein Profiling", *Proteomics* 8, 2211-2219, February 12, 2008.
- [6] Benny Y.M. Fung, Vincent T.Y. Ng, "Meta-classification of Multi-type Cancer Gene Expression Data", 4th Workshop on Data Mining in Bioinformatics, 31-39.
- [7] Herbert pang, Kieta Ebisu, "Analysing Breast Cancer Microarrays from African Americans Using Shrinkage Based Discriminant Analysis", *Human Genomics*, Vol 5, No 1. 5-16, October 2010.
- [8] Junkie Chen, Junzhong GU, "Data Mining Based on Colon Cancer Gene Expression Profiles", *International Conference on Computational and Information Sciences IEEE*, 264-267,2011.
- [9] Jin-Hyuk Hong, Sung-Bae Cho, "The Classification of Cancer Based on DNA Microarray data that Uses Diverse Ensemble Genetic Programming", Seoul 120-749, June 2005.
- [10] Andrew Campen, Yuni Xia, "Mining Gene Expression Database for Primary Human Disease Tissues", 2008.
- [11] Pascale Anderle, Manuel Duval, "Gene Expression Databases and Data Mining", *BioTechniques* 34, S36-S44, March 2003.
- [12] Roberto Ruiz, Jose C. Riquelme, "Incremental Wrapper-based Gene Selection from Microarray Data for Cancer Classification", November 2005.