Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Data Mining in Biodata Analysis

V.Venkata Sai Aditya[1], D. Aruna Kumari[2] and D.Poojitha Bhavana[3]

[1] KLEF University

Received: 13 December 2013 Accepted: 2 January 2014 Published: 15 January 2014

Abstract

For finding interesting patterns in large databases has lot of development in recent years.. Data mining is used in many fields like medicine, securing the data etc. Whereas bio data means the data regarding the biology, medical science, DNA technology and Bioinformatics in-depth analysis. Bio Informatics is the science which can perform managing, finding data, integrating, interrupting information from biological data, genomic, and metadata. Even additional knowledge and complexness can lead to the integration among genes. This paper is all about joining these two fields, the data regarding biology us ing data mining and gives the details of future developments in biodata analysis.

Index terms—

1 Introduction

here are distinct changes in medical research and biodata analysis and there is a lot of growth in medical data collected in medical studies and cancer therapy studies by inventing sequencing patterns, protein-protein interactions gene functions. In biotechnology and bio-data analysis there is a fast growth which has led to the rapid growth in new fields like biodata analysis.

At the same time, according to the recent progress there is a lot of development in the methods of mining interesting patterns and information in large databases, starting from efficient classification methods to clustering, frequent, , serial and structured pattern analysis methods, outlier analysis and visualization. This paper is about how to combine these two fields i.e. data mining and biodata analysis. We need to analyze in which way data mining is helpful in biodata analysis and overview few research problems that may analyze further developments. a) Themes of Biodata Analysis i. Data Cleaning, Data Pre-Processing and Data Integration nature of data there are many challenges in the analysis of medical data. Data cleaning, Data pre-processing and data integration helps in the integration of biomedical data and in the formation of data warehouses for biomedical analysis.

ii. Exploring of Existing Data Mining Tools for Bio Data Analysis

Due to a lot of development in data mining, there are several data mining, machine learning, and applied mathematical analysis systems and tools offered for general data analysis. This analysis is often utilized in biodata analysis and exploration. Data mining analysis is used for biodata analysis including SAS enterprise miner, IBM Intelligent Miner, Microsoft SQLServer 2000, SGI MineSet, and InxightVizServer.Biospecific data analysis systems like GeneSpring, Spot Fire, and VectorNTI can be used in biodata analysis. There are different types of software tools that are developed for resolving the basic bio medical issues. These tools are developing fastly as well..ForBiodata analysis researches should be well trained regarding the usage of tools.

There is much scope for researchers for data mining methods in biodata analysis. Some topics in this view are as follows:

2 b) Similarity in Search and Comparison in Biodata

An essential problem in Biodata analysis is searching similarly and comparing the bio-sequential structures supporting their essential options and functions. For example, the sequences of the genes which are unhealthy and healthy will be compared to notice to note the distinction between the two varieties of genes. This can be

achieved by taking the two categories of genes, then finding the king of factor whether the gene is unhealthy or the healthy one, then comparing the more oftenly occurring patterns of every class. Generally the genetic factor of the disease can be indicated in a way that the diseased sample patterns occur more ofenlyoccuring than the healthy sample patterns. The sequences occurring more frequently in healthy samples indicates the mechanism that protects the body from the diseases. Same type of research can be done on microarray data and protein data to spot the differences in the patterns. Moreover, as the biodata sometimes contains non-perfect matches, it is necessary to develop sequential pattern mining algorithms with in the noisy environment. . Data mining is used in many fields like medicine, securing the data etc. Whereas bio data means the data regarding the biology, medical science, DNA technology and Bioinformatics in-depth analysis. Bio Informatics is the science which can perform managing, finding data, integrating, interrupting information from biological data, genomic, and metadata. Even additional knowledge and complexness can lead to the integration among genes. This paper is all about joining these two fields, the data regarding biology us ing data mining and gives the details of future developments in biodata analysis.

By applying several techniques a variety of bio medical sciences are in use with different geographical dimensions. These are based on data values in bio medical information, genome or proteome databases.Data should be gathered, characterized and clean to extract and analyse information from medicine database and heterogeneous database. The steps for this processing are time taking factors. They need multiple scans for enormous databases to ensure the standards, as a result of he terogeneous and distributed c) Association Analysis There are a lot of studies which has concentrated on the comparing one gene with the other gene. But most of the diseases are no occurred just only by one gene ,it may occur by the combination of two or more genes.Association and correlation analysis strategies can be used to determine the types of genes that may cooccur in final samples. Discovery of groups of such genes or proteins can be done by such analysis. Study of interactions and relationships among the groups and protiens can be done with the help of the analysis done by the association analysis.It is important to develop the serial or structural pattern mining algorithms in the mining environment because the biodata information usually consists of noise or nonperfect matches.

3 d) Cluster Analysis

The process of grouping a group of objects into clusters in which there are similarities in the objects in the same cluster is high and in the objects of different clusters is low is called as clustering.. Clustering is not only in pattern recognition, marketing, social and scientific studies but also in Biodataanalysis.Either Euclidean distances or density are used to determine the algorithms of cluster analysis. The features of biodata analysis are high dimension space, and it is troublesome to review the differentials with scaling and shifting factors in multi-dimensional space and discover the frequently occuring patterns.

4 e) Path Analysis

Complex network among the genes is formed by the biological process. These networks are build ,modeled and visualized using path analysis. The information about biochemical reactions is stored in the database by using the pathway tools. A single genes may not be the reason for causing the disesase,it may be a group of genes responsible for causing a disease process.At the same time there are different stages for different diseases which may become active in any stage of the disease process. The stages of the disease development process will be having a sequence of genetic activities. When this sequence is recognized it will be easy to find the type of the disease for which the future researches can also be developed. By this we can give a better treatment to the diseased people.

5 f) Data Visualization and Visual Data Mining

For aiding the data comprehension the capabilities of human visual systems is used with the help of computer generated representations.. AVS, SGI Explorer, Khoros, MatLab, Visage, SPSS are the general visualization software products. There are many factors for visual data mining and data visualization in the biomedical domain. The first is its huge size.It creates complexities and diversity in biomedical databases. Second, the data producing biotechnologies have been processing rapidly. The demand for biomedical services has been rapidly increasing. Serial patterns of genes are represented by using Graphs, trees, cubes, and chains by different visualization tools.

6 g) Privacy Preserving Mining of Bio-Medical Data

Privacy preserving is the most important factor that any field should have .In biomedical data analysis data regarding genes, proteins, research details should be maintained carefully. For this purpose privacy preserving technique is used. Authorities of hospitals and research institutes will not be able to give the information regarding their hospital details, patient details, their research details etc. Everything should be maintained secretly. Moreover giving such details to other is a crime. So all the details should be secretly maintained. For this purpose privacy preserving should be done with the help of datamining methods which preserves the biomedical data.

7 II.

8 Conclusion

The research frontiers which are data mining and bioinformatics are fast expanding. The research issues in bioinformatics should be examined and the new data mining methods are developed for biodataanalysis which are effective and scalable. There are many methods in data mining which can be used in any field .In biodata analysis data can be preserved, compared, similarities can also be checked using data mining. [1]

Figure 1:

[Baxevanis and Ouellette ()] *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, A Baxevanis , B F F Ouellette . 2001. John Wiley & Sons. (2nd ed.)

[Wang et al. (2002)] 'Clustering by pattern similarity in large data sets'. H Wang , J Yang , W Wang , P S Yu . *SIGMOD'02*, (Madison, WI) June 2002. p. .

[Yang et al. (2002)] 'Mining long sequential patterns in a noisy environment'. J Yang , P S Yu , W Wang , J Han . *SIGMOD'02*, (Madison, WI) June 2002. p. .

[Agrawal and Srikant (2000)] 'Privacy-preserving data mining'. R Agrawal , R Srikant . *SIGMOD'00*, (Dallas, TX) May 2000. p. .

[Hastie et al. ()] *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, T Hastie , R Tibshirani , J Friedman . 2001. New York: Springer-Verlag.