ISSN:0975-9646

Intelligent Data Mining in Autonomous Heterogeneous Distributed Bio Databases

T.Hemalatha, B.G.Gowthami, A.Divya Sri, Ch.Sindhuri, Y.SriDivya, B.BhanuVidhyaKiran

Dept. of Information Science and Technology, K.L.University, Guntur, India.

Abstract- Due to the revolutionary change in data mining and bio-informatics, it is very useful to use data mining techniques to evaluate and analyze bio-medical data. In this paper we purpose a frame work for intelligent data mining system for bio databases especially for diabetic patients depending on their medical test reports. The bio-databases are framed based on the various characteristics involved within the patient suffering with diabetes like: Genetic category, Disease type, country, food habits. System first extracts the relevant, useful, valid and actionable data from bio databases. Bio databases is often autonomous heterogeneous and distributed in nature. The extracted data is preprocessed. After preprocessing, data mining techniques are applied on preprocessed data, local data as well as proprietary data. Then the mined knowledge is integrated with expert system knowledge to assist researchers and scientists in their research work and crucial decision making process.

KEYWORDS: Bio-informatics; Intelligent data ; Autonomous heterogeneous distributed ; Data mining; Expert system; Knowledge base; Diabetes;

I. INTRODUCTION

Diabetes is a disease worrying hundreds of millions of people around the world. In the USA, the population of diabetic patients is about 15.7 million (Breault et al., 2002). It is reported that the direct and indirect cost of diabetes in the USA is $132 billion (Diabetes Facts, 2004). Since there is no method that is able to eradicate diabetes, doctors are striving for ways to fight this doom. Researchers are trying to link the cause of diabetes with patients' lifestyles, inheritance information, age, and so forth in order to get to the root of the problem. Due to the prevalence of a large number of responsible factors and the availability of historical data, data mining tools have been used to generate inference rules on the cause and effect of diabetes as well as to help in knowledge discovery in this area. The goal of this chapter is to explain the different steps involved in mining diabetes data and to show, using case studies, how data mining has been carried out for detection and diagnosis of diabetes in Hong Kong, USA, Poland, and Singapore. On the other hand, recent progress in data mining research has led to the developments of numerous efficient and scalable methods for mining interesting patterns and knowledge in large databases, ranging from efficient classification methods to clustering, outlier analysis, frequent, sequential and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools. The question becomes how to bridge the two fields, *data mining* and *bioinformatics*, for successful data mining in biomedical data. Especially, we should analyze how data mining may help efficient and effective bio-medical data analysis and outline some research problems that may motivate the further developments of powerful data mining tools for biodata analysis.

1.1. Background

Diabetes is a severe metabolic disorder marked by high blood glucose level, excessive urination, and persistent thirst, caused by lack of insulin actions. There are usually three forms of diabetes—Type 1, Type 2, and gestational. It is believed that diabetes is a particularly opportune disease for data mining technology for a number of reasons (Breault, 2001):

- There are many diabetic databases with historic patient information.
- New knowledge about treatment of diabetes can help save money.
- Diabetes can produce terrible complications like blindness, kidney failure, and so forth, so physicians need to know how to identify potential cases quickly.

2. HOW DATA MINING MAY HELP BIODATA ANALYSIS?

Here we list a few interesting themes on data mining that may help bio-data analysis.

2.1 Data cleaning, data preprocessing, and semantic integration of heterogeneous, distributed bio-medical databases.

Due to the highly distributed, uncontrolled generation and use of a wide variety of bio-medical data, data cleaning, data preprocessing, and the semantic integration of such heterogeneous and widely distributed biomedical databases, such as genome databases and proteome databases, have become an important task for systematic and coordinated analysis of bio-medical databases. This has promoted the research and development of integrated data warehouses and distributed federated databases to store and manage the primary and derived bio-medical data, such as genetic data.

2.2 Exploration of existing data mining tools for bio data analysis.

With years of research and developments, there have been many data mining, machine learning, and statistics analysis systems and tools available for use in bio-data exploration and bio-data analysis.

For bio-data analysis, it is important to train researchers to master and explore the power of these well-tested and popularly used data mining tools and packages. A lot of

routine data analysis work can be done using such tools. With sophisticated bio-data analysis tasks, there is much room for research and development of advanced, effective, and scalable data mining methods in bio-data analysis.

2.3. Similarity search and comparison in bio-data.

One of the most important search problems in bio-data analysis is similarity search and comparison among bio-sequences and structures. For example, gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes. This can be done by first retrieving the gene sequences from the two tissue classes, and then finding and comparing the frequently occurring patterns of each class. Usually, sequences occurring more frequently in the diseased samples than in the healthy samples might indicate the genetic factors of the disease; on the other hand, those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. Similar analysis can be performed on microarray data and protein data to identify similar and dissimilar patterns. Moreover, since bio data usually contains noise or non-perfect matches, it is important to develop effective sequential or structural pattern mining algorithms in the noisy environment.

2.4. Association analysis: identification of co-occurring bio-sequences or other correlated patterns.

Currently, many studies have focused on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association and correlation analysis methods can be used to help determine the kinds of genes or proteins that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes or proteins and the study of interactions and relationships among them.

2.5. Frequent pattern-based cluster analysis.

Most cluster analysis algorithms are based on either Euclidean distances or density. However, bio-data often consists of a lot of features which form a high dimension space, and it is crucial to study differentials with scaling and shifting factors in multi-dimensional space and discover pair-wise frequent patterns and cluster bio-data based on such frequent patterns.

2.6. Path analysis: linking genes or proteins to different stages of disease development.

While a group of genes/proteins may contribute to a disease process, different genes/proteins may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible to develop pharmaceutical interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analysis is expected to play an important role in genetic studies.

2.7. Data visualization and visual data mining.

Complex structures and sequencing patterns of genes and proteins are most effectively presented in graphs, trees, cubes, and chains by various kinds of visualization tools. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization and visual data mining therefore play an important role in biomedical data mining.

2.8. Privacy preserving mining of bio-medical data.

Although information exchange is important, hospitals and research institutes may still be reluctant to give out precious bio-medical data due to confidentiality, liability, and other concerns. Thus it is important to develop privacy preserving data mining methods, to maximally protect privacy while achieving effective data mining.

3. SYSTEM ARCHITECTURE

We are going to implement the bio-database in our paper by taking the sample database of the diabetic patients in various parts of the world. By using this biological information of the type of the diabetes the patient is having and various other characteristics, we generate the intelligent mining results to the user. This biological information will be grouped into various databases under a single data warehouse and mining techniques are performed based on the type of the query asked by the user. The bio-databases are framed based on the various characteristics involved within the patients suffering with diabetes. They are:-Genetic category, Patient, Disease, Country, Continent.

3.1. Genetic Category: The genetic features may vary from person to person, persons belonging to one race, persons of one country and persons of one continent but of different countries. So there is necessity to create a database with patients suffering from the disease all over the world.

3.2. Disease Type: There are basically two types of diabetes. They are:-

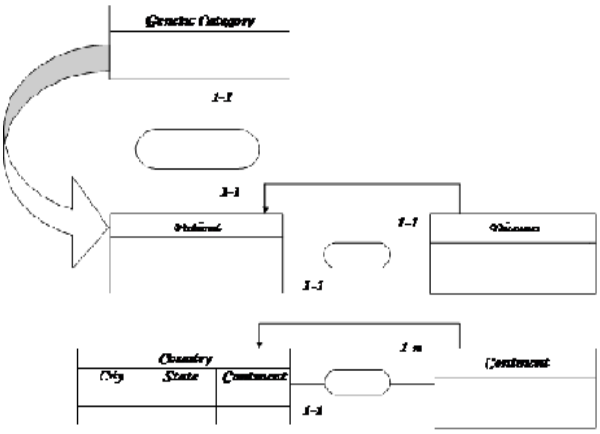
- Diabetes Insipidus and
- Diabetes Mellitus

The patient may be suffering with either of the above diabetic type. So based on the type of the disease present, the characteristics of the genomes differ in these patients. Hence, to distinguish them, a database with the type of the disease is created and maintained.

3.3. Country: The people belonging to various countries may have varied genetic conditions, varied food habits, lifestyle, working conditions and climatic conditions. Hence there will be a similarity among the people belonging to a single country or group living in a certain part of the world. Eg: Most of the Indians suffer from Diabetes Mellitus due to the high intake of carbohydrates like rice. Hence separate data is to maintained by classifying the people according to the country.

3.4. Continent: The people may belong to various countries but may be of one single land, i.e., a single continent. People belonging to a particular continent may also have various similarities in their food habits, climatic conditions, lifestyle etc. So, a database for the people belonging to a continent is maintained and the similarities and dissimilarities are studied during the intelligent mining.

Finally the architecture is as follows:



4. SCREENSHOTS:

4.1. Home page:

4.2.Disease page:

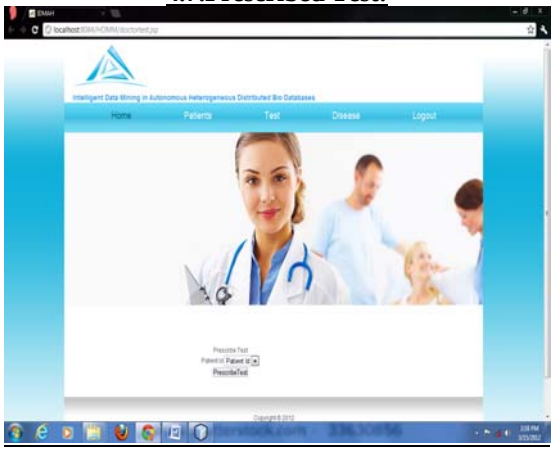
4.3.Genetic Category:

4.4.Disease Classification:

4.5.Patients info:

4.6.Diagonise:

4.7.Prescribed Test:



5. CONCLUSIONS

Both data mining and bioinformatics are fast expanding research frontiers. It is important to examine what are the important research issues in bioinformatics and develop new data mining methods for scalable and effective bio-data analysis. We believe that the active interactions and collaborations between these two fields have just started and a lot of exciting results will appear in the near future.

REFERENCES

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD'00*, pp. 439–450, Dallas, TX, May 2000.

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

[3] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or how to build a data quality browser. In *SIGMOD'02*, pp. 240– 251, Madison, WI, June 2002.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probability Models of Proteins and Nucleric Acids. Cambridge University Press, 1998.

[5] W. J. Ewens and G. R. Grant. Statistical Methods in Bioinformatics: An Introduction. Springer-Verlag, New York, 2001.

[6] J. Han and M. Kamber. *Data Mining:* Concepts and Techniques. Morgan Kaufmann, 2001.

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

[8] A. M. Lesk. Introduction to Bioinformatics. Oxford University Press, 2002.

[9] V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *VLDB'01*, pp. 381– 390, Rome, Italy, Sept. 2001.

[10] H. Wang, J. Yang, W. Wang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *SIGMOD'02*, pp. 418–427, Madison, WI, June 2002.

[11] I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 2001.

T.HEMALATHA received her B-Tech degree in Computer Science and En from Jawaharlal Nehru Technology University, Hyderabad, in 2006, the M-Tech degree in Computer Science from SRM University, Chennai, in 2009. She is working as an Assistant Professor, with Department of Information Science Technology, Koneru Lakshmaiah University, Vijayawada, from 2009 to till now. Her research interests include Cluster technologies, Classification of Data mining techniques.

CH.SINDHURI is pursuing B.Tech in Koneru lakshmaiah college of engineering. She got selected in HCL Company. Her area of interest is Data mining.

B. G. Gowthami is pursuing B.Tech in Koneru lakshmaiah college of engineering. She got selected in TCS Company. Her area of interest is Data mining.

Avirineni DivyaSri is pursuing B.Tech in Koneru Lakshmaiah College of engineering. She got selected in HCL Company. Her area of interest is Data mining.

B. V. Bhanu Kiran is pursuing B.Tech in Koneru Lakshmaiah College of engineering. His area of interest is Data mining.

Y. Sri Divya is pursuing B.Tech in Koneru Lakshmaiah College of engineering. She got selected in HCL Company. Her area of interest is DM