Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Study on Genetic Data and Data Mining -A Way To Future

[1]M.Vijayalakshmi and [2]Dr.V. Vallinayagi,
[1]Assistant Professor, [2]Head, Associate Professor,
[1,2]Department of Computer Science, Sri Sarada College for Woman Tirunelveli-11, India

Abstract: Data Mining is a technique that supports biomedical community and allows new assertions to be made in large volume of Genetic Data sets stored in Genetic Databases. Genetic Databases is a collection of person's personal data relating to the genetic characteristics of an individual concerning health and physiology. Analysing Genetic Data can help to understand the function of human body and reveals new knowledge on human health care. There is 99.5% similarity in the Genetic Pattern among humans. Most human diseases have genetic components such as obesity, hypertension, diabetes, heart diseases, psychiatric problems, and cancers. Analysis of Genetic Data is a growing field, which has the key to the treatment of many diseases. Data Mining techniques can be more efficiently used to organize and analyze the Genetic Data than traditional methods because Genetic databases are complex to analyze. Applying Data Mining techniques on Genetic Data sets can help to improve human health care and also help to reveal health issues like drug reaction and side efforts, etc.In this paper we discuss about the Data Mining applications based on Genetic Data and limitations of applying Data Mining on Genetic Databases.

Keywords: Data Mining, Genetic Data, Genetic Databases, Genetic Pattern, Health Care.

I. INTRODUCTION

Now we are in a Digital World. Every second we receive and provide information through various resources that we are using in our day-to-day life. This information is stored in large databases for further processing whether it is significant or insignificant. So it is unavoidable to think of a process to analyse this information to identify useful information. Data mining is a process of extracting useful information from large databases.Data mining refers to discovery of new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge [1].

Genetic databases are one of the knowledge resources used by Health Care Organizations where Datamining techniques can be applied. Health Care Organizations domain is a use of the disease knowledge management system (KMS) of the hospital case study [2]. Data mining tool is used to explore diseases, operations, and tumour relationships. This tool used to build KMS to support clinical medicine in order to improve treatment quality [2].

II. LITERATURE REVIEW

Tipawan Silwattananusarn and Dr. Kulthida Tuamsuk proposes, data mining applications for knowledge management, and classify the data mining techniques according to the six categories such as classification, regression, clustering, dependency modeling, deviation detection, and summarization [3].

Bradley A.Malin proposes, the biomedical community currently finds itself in the midst of a genomics revolution. Genomic data combined with increasing computational capabilities, provide opportunities for healthcare. The incorporation of genomic data into personal medical records poses many challenges to patient privacy.[5]

Illhoi yoo and others proposes, data mining can help researchers gain both novel and deep insights and can facilitate unprecedented understanding of large biomedical datasets. [6]

III. HEALTH-CARE DATABASES

In general Health-care System domain, the dataset composed of three databases [3]: the health-care providers' database; the out-patient health-care statistics database; and the medical status database [4]. Another data source was from hospital inpatient medical records [2]. Now Genetic databases are used to reveal similarity between individual genetic patterns there by to improve general health.

IV. PREVENTION OF GENETIC DISEASES

The following are the examples of some of the benefits gained by humans through the study of genetic databases.

GENETIC PATTERN: There is 99.5% similarity in the genetic pattern among humans. Chimpanzees are 96% to 98% similar to humans. Cats are 90% gene similarity with humans. Patterns of diseases can have studied in lower animals and the result can be translated in humans because of the genetic similarity. 70% of human genes are found in Zebra fish, have been successfully used to study human genetic diseases.

Fig-1 Henry Krause, a scientist at U of T's Donnelly Centre, is developing zebra fish that carry human DNA segments known as nuclear receptor genes. The transparent fish glow green when a particular organ is hit by a drug being tested. (LUCAS OLENIUK / TORONTO STAR) Website reference: thestar.com.

Genetically modified Zebra fish are used for genetic testing. Their internal organs glow fluorescent green when a

particular organ is hit by a drug being tested. This will help pharmaceutical researchers test new products.

Fig-2 See-through fish offer clear advantage to drug researchers.

GENETIC TEST: There are specific tests that can be done in intrauterine life to know whether the offspring is going to be affected by a genetic disorder. If the foetus is found to have the defective gene the pregnancy can be terminated to avoid the birth of defective children.

GENETIC DELETION: Each community possess a particular genetic pattern paradoxically a loss of particular gene (CCR5) was found to give protection against HIV.

Gene defect can be acquired by genetically manipulating the stem cells in the bone marrow of a HIV-positive person that gives protection against HIV.

For example, the case of "The Berlin Patient" is unique in HIV research. The patient was HIV positive and was stricken by leukaemia. As a part of his cancer treatment, he underwent a bone marrow transplantation. After being treated with large doses of cytotoxic drugs and the radiation to kill cancer cells, he received bone marrow transplanted from a healthy donor.

Fig-3 Timothy Ray Brown, known by many researchers as "the Berlin patient," is the only person to have been cured of an HIV infection

The donor belonged to the small percentage of people who possess the above mentioned co-receptor with defective CCR5 gene. The healthy bone marrow with the defective gene was transplanted into the patient, and the doctors stopped giving him the HIV medication. And the Result? The HIV virus disappeared.

Genetic deletion of a portion of the CCR5 gene leads to protective against HIV.Research is going on whether Knocking out the CCR5 gene can result in a cure for HIV.

V. DATA MINING APPLICATIONS

Data Mining Techniques (DMT) and their applications are classified with respect to the following three areas: Knowledge types, Analysis types, and Architecture types, together with their applications in different research and practical domains. The following are the direction of any future developments in DMT methodologies and applications: (1) DMT is finding increasing applications in expertise orientation and the development of applications for DMT is a problem-oriented domain. (2) It is suggested that different social science methodologies, such as Psychology, Cognitive Science and Human Behaviour might implement DMT, as an alternative to the methodologies already on offer. (3) The ability to continually change and acquire new understanding is a driving force for the application of DMT [7].

Clustering is the grouping together of similar data items into clusters. Clustering analysis is one of the main analytical methods in data mining; the method of clustering algorithm will influence the clustering results directly [9].

VI. THE ROLE OF CLUSTERING

Every day the amount of information being stored on genetic databases is enormous. These databases represent valuable research tools. The genetic databases are too complex and voluminous to be processed and analysed by traditional methods.

Data Mining opens new perspectives where traditional procedures are not adequate for efficient data analysis. Data Mining is a tool that supports research and allows new assertions to be made by disclosing previously undisclosed details in large amount of data.

Clustering is a method used in Data Mining for identifying and describing homogeneous groups of entities, that is clusters, in data sets. The following are the features of clustering method that supports efficient data analysis.

STRUCTURING: Representing data as a set of clusters. Finding homogeneous groups in large amount of dataset. Structuring may be based on humans having same genetic pattern or in same family tree, etc.

DESCRIPTION: Defining clusters in terms of features or different types of phenomenon. Features may be based on personal information, disease affected, genetic test results, response of drug used based on genetic pattern, etc.

ASSOCIATION: Finding interrelation between different aspects of phenomenon by matching descriptions of the same clusters in terms of features related to the aspects. Association helps finding relationship among individuals based on genetic patterns or the phenomenon that is used in cluster formation.

GENERALIZATION: Making general statements about the data structure and, potentially the phenomenon the data relate to. Generalization helps Biomedical Community to take meaningful decision.

VISUALIZATION: Representing cluster structure visually over a well-known Ground Image. For example, Genealogy Tree.

CONCLUSION

Knowledge is an important base for making appropriate Decisions. Discovering the useful knowledge from Knowledge Resources has become a strong demand for development.

The application of DataMining techniqueson Genetic Databases is a growing field which has the key to the treatment of many diseases. The response of diseases to drugs can be decided by the genetic makeup of the individual. In future

prescription of the doctors would be based on the genetic makeup of the individual.

Applying Datamining technique on Genetic/Genealogy Databases would be of immense help to the Biomedical community and Doctors for analysing and treating diseases in future. More advanced techniques are needed to increase the human health care system and the life span of the individual.

References

[1] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases.AI Magazine, 17(3), 37-54.

[2] Hwang, H.G., Chang, I.C., Chen, F.J. & Wu, S.Y. (2008). Investigation of the application of KMS for diseases classifications: A study in a Taiwanese hospital. Expert Systems with Applications, 34(1), 725-733. doi: 10.1016/j.eswa.2006.10.018

[3] Tipawan Silwattananusarn1 and Assoc.Prof. Dr. KulthidaTuamsuk2, Data Mining andIts Applications for Knowledge Management: A Literature Review from 2007 to 2012, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012

[4] Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M. &Kobler, A. (2007). Data mining and visualization for decision support and modeling of public health-care resources. Journal of Biomedical Informatics, 40, 438-447. doi: 10.1016/j.jbi.2006.10.003

[5] Bradley A. Malin, An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future ,JAMIA, VOLUME 12, ISSUE1, JANUARY 2005.

[6] Illhoiyoo and others, data mining in health care and biomedicine: a survey of the literature, A Journal of Medical systems, Aug, 2012, vol 36, issue 4, pp 2431-2448.

[7] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications 39 (2012) 11303–11311

[8] Dr.SankarRajagopal, CUSTOMER DATA CLUSTERING USING DATA MINING TECHNIQUE, International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011

[9] Mythili S, Madhiya E, An Analysis on Clustering Algorithms in Data Mining, IJCSMC, Vol. 3, Issue. 1, January 2014, pg.334 – 340

[10] T. Sajana, C. M. Sheela Rani and K. V. Narayana ,A Survey on Clustering Techniques for Big Data Mining, Indian Journal of Science and Technology, Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, January 2016

[11] Yasodha P, Ananathanarayanan NR. Analyzing Big Data to build knowledge based system for early detection of ovarian cancer. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–7.

[12] Pandove D, Goel S. A comprehensive study on clustering approaches for Big Data mining. IEEE Transactions on Electronics and Communication System; Coimbatore. 2015 Feb 26-27. p. 1333–8.

[13] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005 May; 16(3):645–78.

[14] Yadav C, Wang S, Kumar M. Algorithms and approaches to handle large data sets - A survey. International Journal of Computer Science and Network. 2013; 2(3):1–5.

[15] Bezdek JC, Ehrlich R, Full W. FCM: The Fuzzy C-Means Clustering algorithm. Computers and Geosciences. 1984; 10(2- 3):191–203.

[16] Fahad A, Alshatri N, Tari Z, Alamri A. A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing. 2014 Sep; 2(3):267–79.

[17] Berkhin P. Survey of clustering data mining techniques in grouping multidimensional data. Springer. 2006; 25–71. 12. Macqueen J. Some methods for classification and analysis