ISSN No. 0976-5697

Volume 3, No. 4, July- August 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Quantitative Bio-Medical Data Analysis and Visualization Using Data Mining and Text Mining Approaches

V.V.Jaya Rama Krishnaiah*
Associate Professor,
Department of Computer Science,
A.S.N. Degree College, Tenali, India,
jkvemula@yahoo.com

Dr. K.Ramchand H Rao
Professor
Department of Computer Science and Engineering
A.S.N. Womens Engineering College, Nelapadu India
ramkolasani@yahoo.com

Prof. K. Mrithyunjaya Rao
Director,
Vaagdevi College Of Engineering (M.C.A.),
Warangal, INDIA
kmrkuppa@gmail.com

Abstract: In view of today's information avalanche, recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large databases. The focus of data analysis and data mining tools in biomedical research highlights the current state of research in the key biomedical research areas such as, medical informatics, public health informatics and biomedical imaging. Medicine and biomedical sciences have become data-intensive fields, which, at the same time, enable the application of data-driven approaches and require sophisticated data analysis and data mining methods. Biomedical informatics provides a proper interdisciplinary context to integrate data and knowledge when processing available information, with the aim of giving effective decision-making support in clinics and translational research. Biomedical text data mining is concerned with automated methods for analyzing the content of these documents and discovering and extracting the knowledge in them. Numerical data mining has long been used to uncover patterns in numerical data and make predictions based on those patterns. Text data mining builds on the success of numerical data mining but presents additional challenges. The amount of available biomedical data continues to grow in an exponential rate; however, the impact of utilizing such resources remains minimal. The development of innovative tools to integrate, analyze and mine such data sources is a key step towards achieving larger impact levels. In this Paper, we analyze how data mining may help bio-medical data analysis and outlined some research problems that may motivate the further developments of data mining tools for bio-data analysis and representation of Knowledge.

Keywords: Data Mining, Text Mining, Knowledge Management, Duo Mining, Knowledge Representation, Information Extraction, Information Retrieval

I. INTRODUCTION

The field of biomedical informatics has drawn increasing popularity and attention, and has been growing rapidly over the past two decades. Due to the advances in new molecular, genomic, and biomedical techniques and applications such as genome sequencing, protein identification, medical imaging, and patient medical records, tremendous amounts of biomedical research data are being generated every day. Originating from individual research efforts and clinical practices, these biomedical data are available in hundreds of public and private databases. The digitization of critical medical information such as lab reports, patient records, research papers, and anatomic images has resulted in large amounts of patient care data. Biomedical researchers and practitioners are now facing the "info-glut" problem. Currently, the rate of data accumulation is much faster than the rate of data interpretation [1]. These data need to be effectively organized and analyzed in order to be useful. New computational techniques and information technologies are needed to manage these large repositories of biomedical data and to discover useful patterns and knowledge from them. In particular, knowledge management, data mining, and text mining techniques have been adopted in various successful biomedical applications in recent years[13].

Knowledge management techniques and methodologies have been used to support the storing, retrieving, sharing, and management of multimedia and mission-critical tacit and explicit biomedical knowledge.

Data mining techniques have been used to discover various biological, drug discovery, and patient care knowledge and patterns using selected statistical analyses, machine learning, and neural networks methods.

Text mining techniques have been used to analyze research on electronic patient records. Biomedical entities such as drug names, proteins, genes, and diseases can be automatically extracted from published documents and used to construct gene pathways or to provide mapping into existing medical ontology's [2][3].

In the following sections, we focused the background of knowledge management, data mining, and text mining research on bio medical science. We then analyzed how data mining can be help in data analysis of bio medical data sources.

II. KNOWLEDGE MANAGEMENT, DATA MINING, AND TEXT MINING APPLICATIONS IN BIOMEDICINE

Knowledge management, data mining, and text mining techniques have been applied to different areas of biomedicine, ranging from patient record management to

clinical diagnosis, from hypothesis generation to gene clustering, and from spike signal detection to protein structure prediction. In this section, we focused on some of the relevant research in the field, covering the applications of learning techniques in knowledge management, and data mining and text mining in biomedicine[4].

A. Knowledge Management for Bio Medicine:

One major role of biomedical databases is to serve as a source of vocabulary, i.e., a list of names for the entities represented, strictly speaking, collecting names is the function of terminology.

Besides clinical information, knowledge management has been applied to research articles and reports, mostly via selected information retrieval and digital library techniques. It requires a typical architecture of data organizations, manages the Data extraction, integration and indexing.

B. Data Extraction and Data Integration in bio medical domain:

Data Extraction is a typical mechanism, which extracts the data from the multiple kinds of data sources with reference to the clinical data set. In the biomedical domain, data collection conveying at least one of these two follows kinds of characteristics:

a. They solely focus on highly specialized aspects, like protein interactions, gene expression, and metabolic pathways, but lack general determinants. These include environmental factors, such as behavioral factors, like alcohol consumption, obesity for Heart Decises, Body Mass Weight for Diabetics.
b. They merely provide relatively crude taxonomies that organize entities in a sub-concept hierarchy.

In addition to data selection, Data Integration methods are used for identifying the characteristics of groups obtained through various methods (e.g., the characteristics of patients with respect to the decies symptoms). By integrating the bio medical data, Partesioner may determine the reasons for the typical decises.

C. Data Mining in Bio Medical:

Data mining techniques have been widely used to find new patterns and knowledge from biomedical data.

Because of their predictive power, data mining techniques have been widely used in diagnostic and health care applications. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion.

Data mining is also used to extract rules from health care data the rules generated are similar to those created manually in expert systems and therefore can be easily validated by domain experts. Data mining has also been applied to clinical databases to identify new medical knowledge.

New sequencing technologies and low computation cost have resulted in an overwhelming abundance of biological data that can be accessed easily by researchers. It is not feasible to analyze these data manually, and the gap between the amount of submitted sequence data and related annotations, structures, or expression profiles is rapidly growing. Data mining has begun to play an important role in addressing this problem. Clustering is probably the most widely used data mining technique for biological data. For example, clustering analysis is often applied to microarray gene expression data to identify groups of genes sharing similar expression profiles, various other clustering algorithms like *k*-means clustering (Herwig et al., 1999), fuzzy clustering (Belacel et al., 2004), *k*-nearest neighbors [5] are used.

D. Text mining for Bio medical:

Text mining is also called as intelligent text analysis, text data mining, or knowledge discovery in text uncovers previously invisible patterns in existing resources. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining Text Mining itself is not a function, it combines different functionalities Searching, Information Extraction (IE), Categorization, Summarization, Prioritization, Clustering, Information Monitor and Information Retrieval.

E. Text Mining for Literature and Clinical Records:

Text mining has been widely used to analyze biomedical literature. Because of the large amount of research articles in public databases and the diversity of biomedical research, it is not uncommon that researchers encounter some sequences or new genes that they have no knowledge about. It is quite likely that some important relationships between biological entities remain unnoticed because relevant data are scattered and no researcher has linked them together (Swanson, 1986; Smalheiser and Swanson, 1998). Given the large amount of published literature and that many researchers only specialize in a small sub-domain (e.g., several particular genes), text mining techniques could be invaluable in discovering new knowledge patterns or hypotheses from the large amount of existing and new literature in biomedicine (Yandell and Majoros, 2002). Text mining for biomedical literature often involves two major steps.

a. First, it must identify biomedical entities and concepts of interests from free text using natural language processing techniques. Many text mining algorithms have been applied to this problem. For example, some morphological clues to recognize the heartache like obesity, blood pressure.
b. And then, convert information extracted from the text or unstructured documents into the standardized data set, apply the data mining on data source.

The following Figure 1, shows the typical text Mining Process

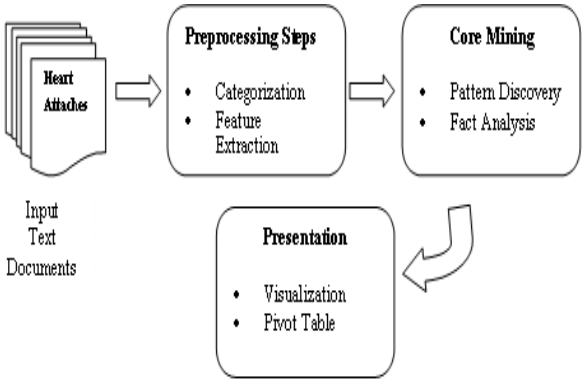

Figure 1: Typical Text Mining Process

III. REASONS FOR SUGGESTING THE DATA MINING AND TEXT MINING FOR BIO MEDICAL DATA ANALYSIS

The question becomes how to bridge the two fields, data mining, Text 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 biomedical analysis. This is the motivation of this talk. Here we list a few interesting themes on data mining that may help bio-data analysis.

A. 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. Data cleaning and data integration methods developed in data mining, such as [3][6][11], will help the integration of bio-medical data and the construction of data warehouses for bio-medical data analysis[12].

B. Exploration of existing data mining tools for biodata 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. Comprehensive surveys and introduction of data mining methods have been compiled into many textbooks such as [7]. There are also many textbooks on bioinformatics, such as [2][8][5][4]. General data mining and data analysis systems have been constructed for such analysis, such as SAS Enterprise Miner, SPSS, SPlus, IBM Intelligent Miner, Microsoft SQLServer 2000, SGI MineSet, and Inxight VizServer. There are also some bio-specific data analysis software systems, such as GeneSpring, Spot Fire, VectorNTI, COMPASS, and SMA (Statistics for Microarray Analysis) in R. These tools are evolving as well. 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.

C. 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 biomedical data usually contains noise or non-perfect matches, it is important to develop effective sequential or structural pattern mining algorithms in the noisy environment [9].

D. 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.

E. 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 [10].

F. 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 [14].

Finally, with this paper, we propose a new architecture for Bio Medical data analysis as shown in the Figure 2

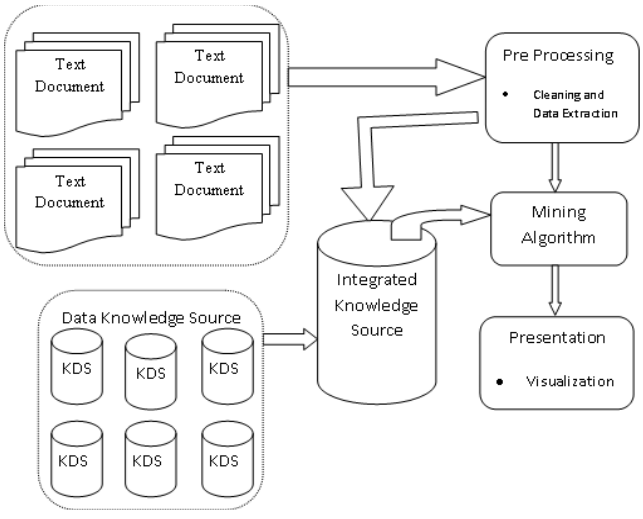

Figure 2: Typical Architecture proposed for Bio Medical data Analysis

IV. KNOWLEDGE REPRESENTATION

Knowledge Representation phase is very important level in Bio Medical Data Analysis. It concerns about the different facts and rules about a subject. It represents the abstract of the knowledge.

Duo mining tools were used for making and representing data. Duo Mining is the combination of Text and Data mining Processes.

Different kinds of approaches have been employed for representing facts like

a. Histograms: *a* ***histogram*** is a graphical representation showing a visual impression of the distribution of data. This is constructed with intervals of the facts. For Example no of people effected with diabeties with respect to BMI between age group 0-20

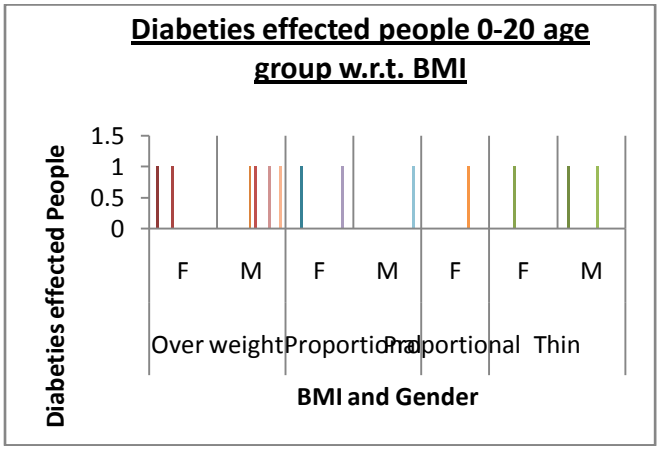

Figure 3: Histogram

b. Pie Charts: It is a circular chart divided into sectors each of whose length is proportional the quantity it represents. For Example, Finding out the number of people having the Good BMI and corresponding Cholesterol levels.

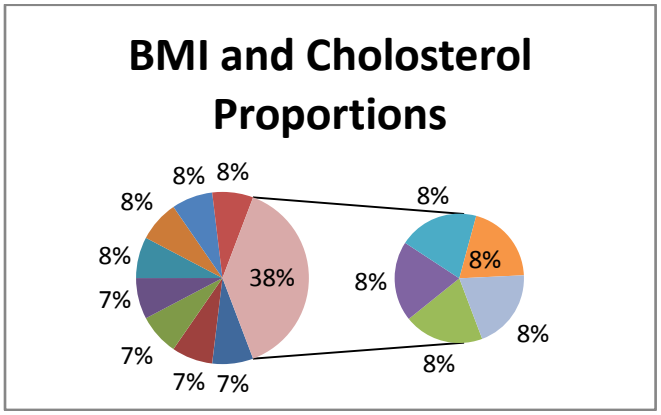

Figure 4: Pie Chart

c. Line Chart*:* A line chart or line graph is a type of chart which displays information as a series of data points connected by straight line Segments. It is a basic type of chart common in many fields. A line chart is often used to visualize a trend in data over intervals of time – a time series thus the line is often drawn chronologically. For Example comparing the different run performance elevation of different methods in detecting the diabetes.

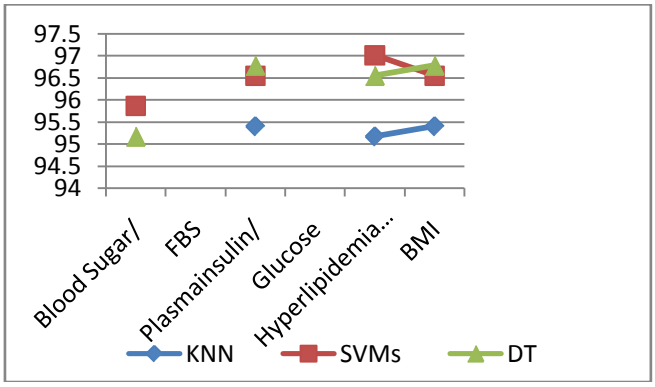

Figure 5: Line Chart

V. CONCLUSION

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. The proposed architecture for Bio Medical Data Analysis can help in integrating both text and data sources for better analysis of symptoms for a deices. It eliminates the extra burden on data analysis because it converts unstructured data into supervised or structured fashion as maintained in Integrated Knowledge Source.

VI. ACKNOWLEDGMENT

Our thanks to Department of Computer Science and Engineering; Acharya Nagarjuna University; Department of Computer Sceince, Annabathuni Satyanaraya Degree College, Tenali, for providing necessary facilities to carryout this Review Paper. My Sincere thanks to my friend, D.V. Chandra Sekar, Associate Professor, Computer Sceince, T.J.P.S College, Guntur, who helps me making of this Paper.

VII. REFERENCES

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

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

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

[4]. 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.

[5]. Jiawei Han, How Can Data Mining Help BioData Analysis?

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

[7]. R. Agrawal and R. Srikant. Privacy-preserving data mining. In SIGMOD'00, pp. 439–450, Dallas, TX, May2000.

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

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

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

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

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

[13]. J. Yang, P. S. Yu, W. Wang, and J. Han. Mining long sequential patterns in a noisy environment. SIGMOD' 02, pp. 406–417, Madison, WI, June 2002.

[14]. Apte, C. Data Mining An Industrial Research Perspective. IEEE Computational Science and Engineering. v 4,1997

Short Bio Data for the Author's

V.V.Jaya Rama Krishnaiah received Master's degree in Computer Application from Acharya Nagrajuna University,Guntur, India, Master of Philosophy from Vinayaka University, Salem . He is currently working as Associate Professor, in the Department of Computer Science, A.S.N. Degree College, Tenali, which is affiliated to Acharya Nagarjuna University. He has 13 years teaching experience. He is currently pursuing Ph.D., at Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. His research area is Clustering in Databases. He has published several papers in National & International Journals.

Dr.K Ramchand H Rao received Doctorate in from Acharya Nagarjuna University, Master's degree in Technology with Computer Science from Dr. M.G.R University, Chennai, Tamilnadu, India. He is currently working as Professor and Head of the Department, Department of Computer Science and Engineering, A.S.N. Women's Engineering College, Tenali, which is affiliated to JNTU Kakinada. He has 18 years teaching experience and 2 years of Industry experience at Morgan Stanly, USA as Software Analyst. His research area is Software Engineering. He has published several papers in National & International Journals.