ISSN: 2321-7782 (Online)
Volume 3, Issue 10, October 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Bio-Informatics using Data Mining Techniques

V. Saranya
Research Scholar,
Department of Computer Science,
H.H. The Rajahs College (Autonomous), India

Abstract: ***This article highlights some of the basic concept of bio-informatics and data mining techniques. Bio-informatics, an upcoming field in today's world, in which involves use of large amount of databases can be effectively searched in through data mining techniques to derive useful rules. Data mining is especially used in microarray analysis in which is used to study of the activity of different cells and under different conditions. Data mining techniques are compares two algorithm of association rule mining. The explosive growth of biological information generated by the scientific community and all over the world has led to storage of voluminous data. It also highlights some of the current challenges and opportunities of data mining in bio-informatics.***

Keywords: Association Rule Mining, Bio-informatics, Data Mining, Genetic Algorithm using data mining, Microarray on data mining, Soft Computing Techniques.

I. INTRODUCTION

In modern world, quick developments in genomics and proteomics have generated a large amount of biological data storage in use of data mining techniques. The important of this new field of inquiry will grow as we continue to generate and integrate a large amount of genomic and proteomic and other data [3]. Bioinformatics is the science of managing, mining, integrating and interpreting of information from biological data at the genomic and proteomic and phylogenetic, cellular or whole organism levels. A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problem. Different type of bio-informatics problem to solve used in data mining techniques on microarray. Example of this type of analysis include protein structure prediction, gene classification and cancer classification based on microarray data analysis, cluster of gene expression data analysis, statistical modeling of protein to protein interaction.etc. Data sets representing entire genomes' worth of DNA sequence, such as those produced by the Human Genome Project- [13], are difficult to use without observations, which label the locations of genes and regulatory elements on each chromosome. Drawing conclusions from these data sophisticated computational analysis. Therefore, we use see a great potential to increase the interaction between data mining and bio-informatics. This purpose of this paper is to provide an overall understanding of data mining and soft computing techniques and their application and usage in bio-informatics.

II. LITERATURE REVIEW

We review fresh result in literature data mining for biology and discuss the need and the steps for a challenges and evaluation for this field. Literature review of data mining has advanced from simple recognition of terms to extraction of interaction relationships from complex sentences and has broadened from recognition of protein-protein interaction to a range of problems such as improving homology search identifying cellular place, and so on. To encourage the participation and accelerate improvement in this expanding field, we propose creating and challenges, evaluation, and we describe two specific applications in this context. Over the year, many of the basic problems in bio-informatics, such as protein structure prediction on gene decision and data retrieval and integration are motionless open. In recent year, high-throughput experimental methods in molecular biology have resulted in huge amounts of data. Mining bio-informatics data is an emerging area of intersection

between bio-informatics and data mining techniques. The purpose of this book is to facilitate collaboration between data mining researchers and bio-informatics.

III. BIO INFORMATICS

The term bio-informatics was coined by Paulien Hogeweg. It was first used since late 1980's has been in genomics and genetics, particularly in that field of genomics and proteomics involving in large-scale of DNA sequencing. DNA abbreviation of Deoxyribo Nucleic Acid.

Some of the particular area of research in bio-informatics includes:

- Analysis of gene expression in DNA Sequence
- Analysis of alterations in cancer
- Analysis of protein expression in DNA sequence
- Comparative of genomics
- Sequence of analysis
- High-throughput in image analysis
- Modeling of biological systems
- Protein structure prediction
- Protein-Protein docking

A. Purpose and use of bio-informatics:

Bio-informatics is used in analyze of genomics, proteomics, protein sequences and three-dimensional modeling of bio-molecular and biologic systems, etc.

- Comparison and alignment of DNA, RNA and protein sequences.
- Gene regulatory of network identification process.
- Interpretation of gene sequences expression and microarray data.
- Molecular plan and molecular docking.

Therefore the aims of bio-informatics are:

- To use of biological data to analyze and interpret of the result in a biological meaningful manner.
- To organize the data in a way that allows researchers to create and access information.

To develop the tools that facilitates the analysis and management of data.

B. Bio-informatics tools:

Bio-informatics used in different research area and different tools in reference the details.

Bio-informatics research area	Tool(Application)
Sequence Alignment	BLAST, CA-BLAST, HMMER, FASTA
Multiple Sequence Alignment	MSA Probs, DNA Alignment, MultiAlin, DiAlin
Gene Finding	Gen Scan, Genome Scan, Gene Mark
Protein Domain Analysis	Pfam, Blocks, Prodom
Patteren Identification	Gibbs Sampler, AlignACF, MEME

Genomic Analysis	SLAM, Multiz

IV. METHODOLOGY USED

4.1 Data mining:

Data mining are refers to extracting or "mining" knowledge from large amount of data. Data mining is also sometimes called knowledge Discovery in Databases (KDD).

Researchers are identified two important goals of data mining techniques Prediction and description.

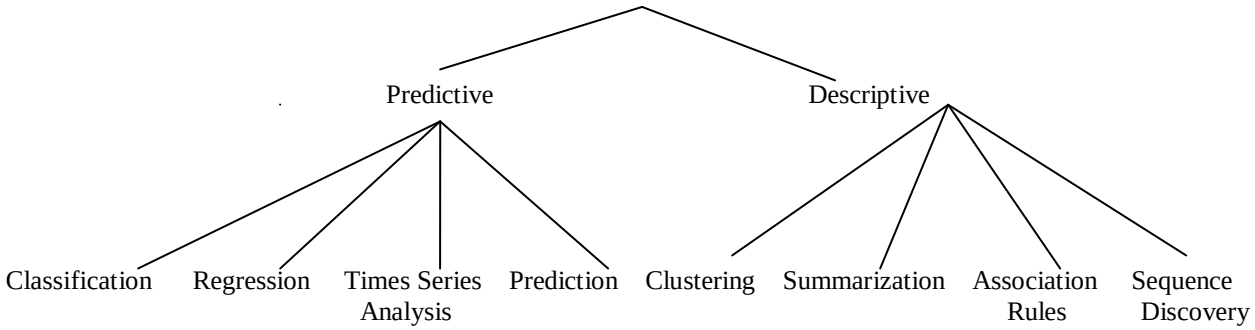

Figure: 1 The data mining techniques two "high-level primary goals of data mining, in practice are prediction and description techniques.

Prediction makes use of presented the variables in the database on order to predict the future values of interest and description of focuses on finding patterns describing the data and the subsequent of presentation for the user explanation[14].

Knowledge discovery in databases as a process is depicted in consists of an iterative sequence of the following steps:

- Selection: Acquire data from various sources.
- Preprocessing: Clean data.
- Transformation: Convert old to new format Transform old to new format.
- Data Mining: Obtain desired results.

Interpretation/Evaluation: Current result to user in meaningful manner.

4.1.1. Association Rule Mining:

Association rule mining or market basket analysis. Association rule mining searches interesting correlations among the items in given dataset [4]. Association rules mining techniques has a large number of application other than market basket analysis, including application in marketing, customer segmentation, medicine, electronic commerce, and e-business, classification, clustering, web mining, bio-informatics and finance. Let IT = (it1, it2, ……., itn) be a set of literals called items. Let D be a set o transaction where each transaction T is a set of items such T≤IT. Let P is a set of items. A transaction T is said to contain p if and only if P≤T. An association rule is an implication of the P=>Q where P<IT, Q<IT and P∩IT=Ø. The rule P=>Q hold in the transaction set D with confidence C if C% of transactions in D that contain P also contain Q. The rule P=>Q has support S in the transaction set D if S% of D contains P Q.

Support:

Support (P) = (Number of times P appears)/N=P(P)

Support (PQ) = (Number of times P and Q appear together)/N = P (P∩Q)

Confidence:

Confidence = Support (PQ)/Support (P) = P (P∩Q)/P (P) =P (Q/P)

Problems:

(1). Discover the item set that have transaction support above a pre-determined minimum support.

(2). Use the large item set to generate the desired association rules for the database.

We therefore stress on two association algorithm:

A priori and partition algorithms and then compare the two algorithms.

(A) Apriori:

The basic algorithm for decision the [5] association rules was first proposed in 1993. In1994, an improved the algorithm was proposed. Our discussion is based on the algorithm called the apriori algorithm. Apriori uses a "bottom up" where the frequent subsets are extended one item at a time (a step known as candidate generation) and groups of candidates are tested against the data.

The phonetic signs used in the algorithm are:

1. L k: set of large k-item sets (those with minimum support of apriori)
2. C k: set of candidates k-item sets (potentially large item sets)

Example

TID	Items
A	10,20,50
B	10,50,30,40
C	10,40,30

Table: 1 Data Item set

TID	ITEMS				
	10	20	30	40	50
A	1	1	0	0	1
B	1	0	1	1	1
C	1	0	1	1	0

Table: 2 Storage database

Partitioning Algorithm:

The partition algorithm is based on the observation that the frequent sets are normally very few in number compared. It reduces the number of database scans to 2. I t reduces the time delay and the repeated data scan avoid the algorithm. It divides the database into small partitions such that each partition can be handled in the main memory. Let the non-overlapping partitions of the database be D1, D2, …, D p.

In the first scan, it finds the locally frequent item sets in each partitions Di ($1 \le i \le P$), $\{P \mid P.\ count \ge s \times \mid Di \mid\}$. Since each partition can fit in the main memory and there will be no extra disk I/O for each partition after loading the partition into the main memory [6]. In the second scan, it uses the property that a frequent item set in the whole database must be locally frequent in at least one partition of the database. Then the unions of the locally frequent item sets found in each partition are used as the candidates and are counted through the whole database to find all frequent item sets. However for a skewed data distribution must of the item sets on the second scan may turn out to be small and thus destroying a lot of CPU time counting false item sets.

V. PROPOSED WORK

Data mining is especially used in microarray analysis which is used to study the activity of different cells under different conditions. Two algorithms under association rule mining techniques were implemented for a large database and compared with each other.

Association Rule Mining: (a) apriori (b) partition

Owing to the involvement of large datasets and the need to drive results from then, data mining techniques can be effectively put in use in the field of bio-informatics. The techniques can be applied to find associations among the genes. Further, data mining techniques can be made more efficient by applying genetic algorithms which greatly improves the search and reduces the execution time.

VI. Experimental Results

Number of candidate set generated algorithm

Apriori-49

Partition-90

Number of database scan algorithm

Apriori-3

Partition-2

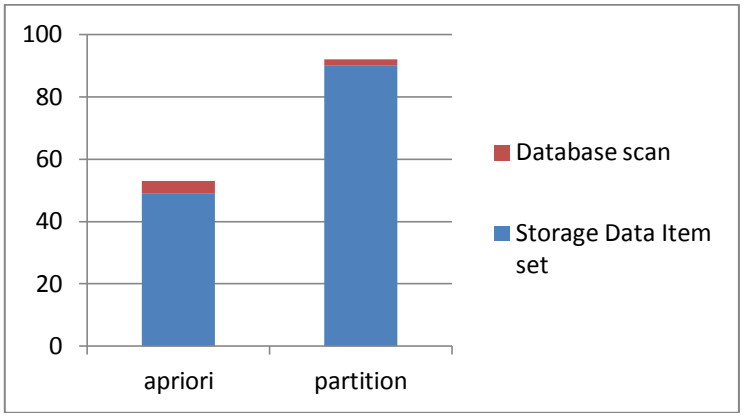

Figure 2: Compare the apriori and partition algorithm

The experimental result show that apriori works better than partition algorithm for smaller database while partition works better for larger databases partition require less I/O overhead then apriori because it requires lesser database scans [7]. Compare the two algorithm of association rule mining technique. In the experimental result in best algorithm of partition in a large data storage and reduced database scan less and time save and storage the valuminous data.

VII. Data Mining Application

It may be used in applied successfully in many areas from business to science and other than data.

1. **Science Applications:** It may be used in astronomy, bio-informatics, drug discovery and many more.
2. **Business Application:** Data mining has been used in advertising, CRM (Customer Relationship Management), investment, manufacturing, sports and entertainment, telecom, e-commerce, e-business, targeted marketing, healthcare, etc
3. **Other Application**: Large number of organizations now employ data mining as a secreate weapon to keep in pace or gain a competitive edge.

VIII. Soft Computing Technique

Soft computing techniques refers to a collection of computational techniques in computer science and artificial intelligence, machine learning and some engineering disciplines, which try to study, of model and analyze very complex phenomena [15].The complementary of fuzzy logic, neural networks, genetic algorithm and probabilistic reasoning has an important consequence in many cases.

Fuzzy Logic:

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than exactly deduced from classical predicate logic. Fuzzy logic can be used to control household appliances such as washing machines (which sense load size and detergent concentration and adjust their wash cycle according) and refrigerators.

Neural Network:

Artificial neural networks are computer models built to emulate the human pattern recognition function through a similar parallel processing structure of multiple inputs. A neural network consists of a essential processing elements (also called neurons) that are distributed in a few hierarchical layer.

Genetic Algorithms:

Genetic algorithms have found a wide gamut of applications in data mining, where knowledge is mined from large databases. This concept of survival of the equipments proposed by Darwin is the main reason for the robust performance of Gas. As a simple example, suppose that samples in a give training set are described by two Boolean attribute A, and B and that there are two Class C and D. The rule \IF A and not B THEN D" can be encoded as the bit string \100", where the two leftmost bits represent attributes A and B respectively and the rightmost bit represent the class. The rule \IF not A and not B THEN C" can be encoded as string\001". If an attribute has n values where n>2, then n bits may be used to encode the attribute's values.

Only four different types of nucleotides (or bases) are used in DNA molecules:

Adenine, Guanine, Cytosine, and Thymine (A, G, C and T [ ]). Different combinations of only 20 various amino acids are used to build all of the proteins in a human being [10][11].

A. Microarray:

Microarray data pose a great challenge for computational techniques, because of their large dimensionality (up to several tens of thousands of genes) and their small sample size-[8].

Furthermore, additional experimental complications like noise and variability give the analysis of microarray data an exciting domain. DNA microarray- [9] are created by robotic that arrange minuscule amount hundreds or thousands of gene sequences an a single microscope slide.

Researchers that have a database of over 40,000 gene sequences expression that they can use for this purpose. This expression pattern is then compared to the expression pattern of a gene responsible for a disease.

Use of Microarray:

- ❖ Gene Discovery
  - Tissues profiles
  - Time course data
  - Altered genetic backgrounds
- ❖ Comparing tissues/genotypes

THE WORK OF MICROARRAY

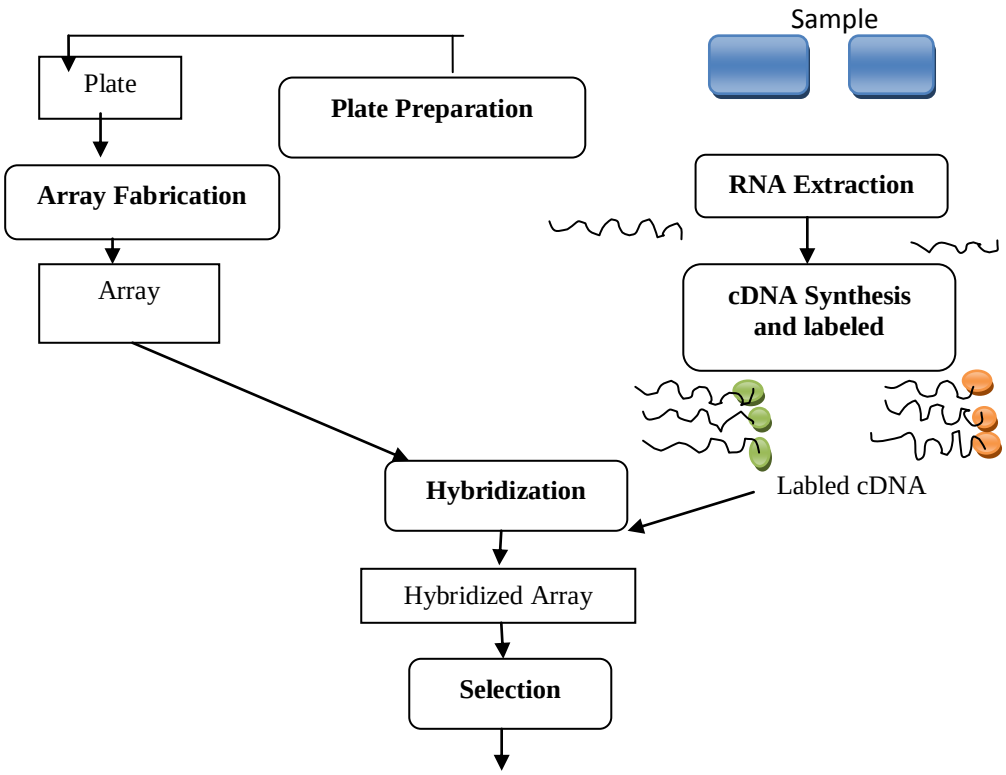

Figure 3: Work of Microarray analysis

IX. CONCLUSION AND FUTURE WORK

In this article report an in-depth study of varied bio-informatics and data mining techniques was made. The report then gives an introduction to molecular biology and bio-informatics. Soft computing techniques are closely used in the field of bio-informatics to solve difficult problems. The can be used for drug treatment considering the response of the genes to drugs paving the way for diagnosis of incurable diseases like AIDS, Alzheimer's disease. Data mining and bio-informatics are fast growing research area today. It can be also used to identify the mechanism underlying the biological processes such as growth of ageing and to track the process of our evolution. Data mining technique is therefore a helpful to solve the problem of huge data faced by researchers in their search to solve the confuse of our life.

References

1. Pujari, Arun Data mining Techniques, Nancy: Universities press,2001.
2. Han, Jiawei and Kamber , Micheline. Data mining: concepts and techniques, San Francisco: Morgan Kaufmann publishers Inc., ca, 2000.
3. Lee, Kyoungrim (2008).computional study for protein-protein Docking using Global optimization and Empirical potentials, Int.J.Mol.Sci
4. Agarwal,R.,Imielinski,T. and Swami, A."Mining associations between sets of items in massive databases ACM SIGMOD International conference on Management of Data, pp 207-216, Washington DC, May 1993.
5. Hipp, Jochen, Guntzer, Ullrich and Nakhaeizadeh, Gholamreza,"Algorithms for Association Rule Mining-A general survey and comparison".SIGKDD explorations, vol2,Issue-1,pp58-63.mar-2004.
6. Agarwal', Rakesh and Srikant, Ramakrishnan."Fast Algorithms for Mining Association Rules in Large Databases", proceedings of the 20th International conference on Very Large DataBases, pp.487-499, septemper12-15, 1994.
7. Gyorodi, Robert S."A comparative Study of Interative Algorithm in Association Rules Mining", studies in Informatics and contaol, vol-12,no-3, pp205-212, sept 2003.
8. Piatesky- Shapiro, G.and Tamayo,P:"Microarray Data Mining:Facing the Challenges."SIGKDD Explorations 5(2),pp 1-5, 2003.
9. Liu, lL., Yang,Jiong.and Tung, Anthony."Data Mining Techniques for Microarray Datasets." Proceedings of the 21th International conference on data 13 Engineering 2005 IEEE. Pp182-192,2005.
10. Shah, Shital c.and Kusiak, Andrew,"Data Mining and Genetic algorithm Based Gene Selection",Artificial Intelligence in Medicine 2004,(31),pp183-196 vol-2139,2004.
11. Goldberg, David E.Genetic Algorithm in search, optimization and mechine learning. Boston:Addison-wesley Longman publishing Co.,1989.
12. Bergeron, Bryan.Bioinformatics computing New Delhi: pearson Education, 2003..

13. Gilbert, D.(2004). Bioinformatics software resources. Briefings in Bioinformatics.

14. Aluru, S.,ed.(2006) Handbook of Computational molecular Biology. Champman & Hall/Crc.

15. Mickey sahu and Ashish Shrinivastava (2013),Mobile Learing-Telecommunication Infrastucture and usage in Rular and Remote Areas Students: A Review, Indian journal of applied research.

