Data Mining in Bioinformatics: Study & Survey

Saliha V S
St. Joseph's college
Irinjalakuda

***Abstract*--Large amounts of data are generated in medical research. A biological database consists of a collection of data and information about different biological aspects. The information in these databases can be searched, compared, retrieved and analyzed. Best way to analyze the biological data is data mining. Mining on biological data has great importance in today's world. By the size and complexity, getting of data from biological databases is a complicated process. Biological databases have become important tools in assisting scientists to understand several biological phenomena. This article explores data mining in bioinformatics from the perspective of the process of discovering biological data. The basic data mining process on biological data requires knowledge discovery as the first step. This article reviews the numerous technologies that can be applied to support data mining. Text mining focuses on the importance of mining biomedical literature for data on functions to complement the sequence and structure data mined from nucleotide and protein databases. General purpose and bioinformatics tools are available for data mining.**

I. INTRODUCTION

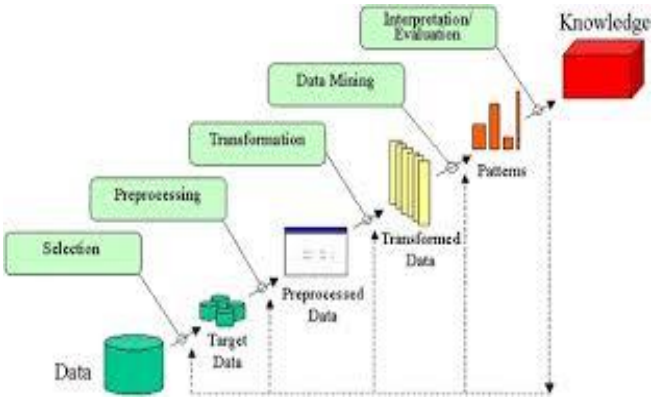
Bioinformatics is rooted strongly in life sciences and also with the computer based information sciences. Bioinformatics is information technology dealing with the maintenance and use of data in molecular biology using computers. In short, it is the information technology applied to molecular biology. It is also called computational molecular biology. Rapid research in different fields of molecular biology has generated a vast enormous amount of data that can't be handled manually. Eg: data from genome mapping, sequencing, etc. Bioinformatics consists of the collection, maintenance, distribution, analysis and usage of the large amount of data generated in molecular biology for biological investigations. Suitable algorithms are created for processing that data using computers to resolve problems in analyzing the experimental data in molecular biology. To store and retrieve the bibliographic or biological information from the databases and to analyze the sequence patterns and to extract the biological knowledge of the sequences, effective and efficient tools are needed. Data mining is the process of knowledge discovery from a huge amount of data that go beyond simple analysis. It is also known as knowledge discovery in data. It discovers previously unknown patterns.

II. DATA MINING

Preceding processes are modified in data mining to support new hypothesis suggested by the data which takes place iteratively. The following steps involves in knowledge discovery process:

- selection and sampling
- preprocessing and cleaning
- transformation and reduction
- data mining
- evaluation
- visualization
- designing new queries

The relative timing of sequences in the knowledge discovery process varies depending on whether the source of data is a data warehouse or one or more separate databases. A controlled vocabulary, usually implemented as part of a data dictionary in which single word can be used to express a given concept required in data mining. Knowledge discovery process is an iterative process which takes feedback from each stage.



A. Selection and sampling:

Data mining takes place on data warehouses or one or more of the biological databases. In a large amount of data it is not mean that it takes all possible relationship. It selects only few samples to evaluate the relationships and can be used to determine which data should be mined further using the complete data warehouse.

B. Preprocessing and cleaning:

The knowledge discovery associated with the bulk of work is that to prepare data in association with data mining for the actual analysis. Preparatory activities are also performed to some extend in the stage of data warehouse creation and that includes

- Data characterization: It creates a high-level description of the nature and content of the data to be mined.

- Consistency analysis: It is the process of determining the variability in the data which is independent of the domain.
- Domain analysis: Validation of the data values in the larger context of the biology is called domain analysis. This requires a person who is familiar with biology for the creation of the heuristics that are applicable to the data.
- Data enrichment: It minimizes the limitation of single data source by drawing from multiple data sources.
- Frequency and distribution analysis: Depending on the frequency of occurrence, it place weights on values.
- Normalization: This process uses predefined range of values to transform the data values from one representation to another. Absolute, rank, ratio, etc. are the most common scales used in normalization.
- Missing value analysis: The process which is the final preprocessing and cleaning activity that detects, characterizes and deals with missing data value. It uses mean, mode or median of the relevant data to substitute the missed data.

C. Transformation and reduction: Through sampling or summary statistics, this phase reduces the minimum size of the datasets possibly. The transformation does not allow the data from multiple sources and it directly supports the data mining and knowledge discovery process.
Data mining methods: Data mining extract pattern using classification, regression, link analysis, segmentation or deviation detection.

- Regression method assigns data to a continuous numerical variable based on statistical methods. It used to extrapolate trends from a few samples of the data.
- Link analysis is the evaluation of apparent connection or links between data in the database or data warehouse.
- Segmentation is process of identification of the classes or groups according to some metric that behave similarly.
- Deviation detection involve with the identification of the data values that are outside of the norm using definition of existing model or evaluation of the ordering of observations.

D. Evaluation:
This phase includes the interpretation of the patterns that are identified by the data mining analysis.
Visualization: It's an optimal stage and can range from the conversion of tabular listings of data summaries to pie charts and similar business graphics.

E. Designing new queries:
The iterative process, data mining suggests new hypothesis or hypothesis originates from other research. This phase involves with the formulation of new queries and revisiting of the selection and sampling stage of knowledge discovery process.

III. MACHINE LEARNING

Machine learning techniques are used to perform pattern matching and pattern discovery. It comprises of different methods that are converged to artificial intelligence, psychology, statistics, adaptive control theory and biological modeling.

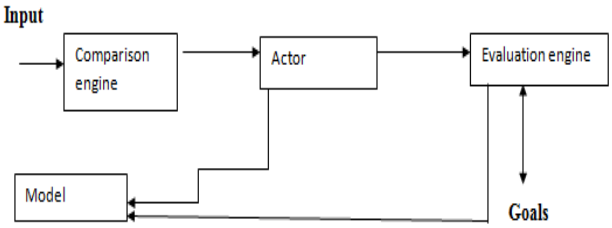


Figure: Machine-learning process.

Figure shows the general machine learning process regardless of the underlying technology. The comparison engine compares the input data with the underlying structure and the result is fed in to the software actor that initiates some changes to the data. The output from actor is evaluated by evaluation engine with underlying goals of a system as a reference. In the model the changes is directed by the feedback from actor and evaluation engine. The output is the pattern associated with the input data. Machine learning are of two types: supervised learning in which system is trained with set of examples, called the training set and each input has its own specific output. In unsupervised learning there is no specific output associated with the given input.

Machine Learning Technologies	Classification	regression	segmentation	Link analysis	Deviation detection	
Inductive logic programming	X	X				
Genetic algorithms	X	X	X			
Neural networks	X	X	X			
Statistical methods	X	X	X		X	X
Decision trees	X		X			
Hidden Markov models	X					

Table: Machine learning technology and their applicability to data mining methods.

- Inductive logic programming: It categorizes the data using a set of rules or heuristics. Inductive logic programming derives hypothesis using background knowledge and a set of examples.

[7]Inductive logic programming is a covering algorithm.

theory = empty
While (positives left uncovered):
- construct a rule
- theory=theory + rule
- remove positives covered by rule

return theory

- Genetic algorithms: The basic Darwinian approach 'survival of the fittest' is used by genetic algorithm. It allows algorithm to adapt to the needed environment dynamically by continuously iterative the process as indefinite. It generates large amount of solutions randomly to solve a problem. The optimization problem that describes chromosome encoding can be solved by genetic algorithms. The optimization response time is not constant in genetic algorithms. It works through following steps:
  - Creates population of strings
  - Evaluates each string
  - Selects the best string
  - Creates new population of strings by genetic manipulation
- Neural networks: A neural networks consists of a set neurons which is the basic processing elements and in a few hierarchical layers, they are distributed. It learns the output patterns in association with input patterns. The learning process categorizes new patterns and trends from data are extrapolated. Neural network represents patterns as input nodes and it produces the output pattern. Working of a neural network is independent of its problem domain. Neural network uses rule-based expert systems which have rules in human-readable form:

  IF condition THEN goal

  Validation of inner workings of neural network is difficult. A neural network classifier is trained before its usage.
- Statistical methods: To support data mining statistical methods are used in feature extraction, clustering or classification. In feature extraction using statistical methods, it defines the data attribute that may be unclear or difficult to understand such as noise in the data, imperfect measurement or improper data processing. A variety of pattern classification methods using the concept in statistics can be applied to data mining. Statistical data mining methods describes complex patterns in terms of simpler patterns based on structural pattern recognition. Predictive modeling predicts the missing data by using data in the database based on regression or classification. The similar data are grouped into subsets (data segmentation), called cluster analysis. It can express similarities and differences by constructing tree diagram using abundant data about number of objects. Cluster analysis uses model-, partition- and metric based methods.
- Decision trees: It represents questions and answers arranged hierarchically which leads to classification. Relative occurrence frequencies representation is not possible in decision trees. Decision tree takes advantage of this in some cases. For example multiple samples from the same or closely related species may skew the relative abundance of some properties over other in classification of globins from a variety of species. It is a predictive model approach. A decision tree in which target variable that takes finite set of values, known as classification tree. Decision trees that take continuous values for target variable are called regression trees.
- Hidden Markov models: It is a statistical model that constructs classifiers. It acts as state machine for an ordered sequence of symbols in which a symbol generates a transition made from one state to next state. Transition probability specifies the transition between states. A process moving from one state to another depending on 'n' previous states is called Markov process. If the choice of next state is affected by 'n' states, then the process is called an order n model. In Markov chain, the probability for the transition from one state to next not varies with time. A Markov chain in which states in chain are hidden called Hidden Markov model (HMM). HMM must be trained before its usage like neural network classifier.

IV. TEXT MINING

A lot of information is available in online as a form of unstructured free text. Extracting the data from the online document in several languages is a difficult task. The information of various fields of bioinformatics such as genomics, proteomics, pharmacogenomics, etc are available in online such as in Pub Med, PLMITRNA, etc. Text mining is complicated due to the variation of representation of data in a text document.

Natural language processing (NLP) is a technology that contains various computational methods from simple word extraction to semantic analysis. In NLP the keyword is first extracted and the document that contains the keyword is copied to a local database for the further review. For keyword extraction NLP uses statistical method. The first step is named entity recognition is the process of identifying all the instance of the names within a collection of text. This involves recognizing biological entities for future extraction of relationships. The next step is the text classification that identifies all the document or part of the document that contains the keyword. Some biological entities have more than one name or abbreviations. Sometimes we use this abbreviation for convenience. By resolving abbreviations of the word or replacing synonyms and acronyms to the vocabulary that controlled, finishes the processing part. Heuristics, grammars and statistics are used in the analysis phase of NLP. Heuristics model use

knowledge based rules for analysis. Mathematical model are used to get the meaning and context of the words in statistical method. Language models are used in grammar based analysis for deriving information from the text that processed.

V. CONCLUSION

Bioinformatics and data mining are fast expanding research areas. Biomedical research contributes a lot of biological data and it's available in various resources. Data mining which is one of the phases in the knowledge discovery process simplifies the difficulties for the analysis of biomedical research. Data mining finds relationship between data and discovers new patterns from existing data. Data mining application is helpful in cancer detection, drug design, simulation for health care issues, etc. In a future with improved and modified data mining methods simplify the bio data analysis. The tools for the text mining are not part of biomedical research area. The huge size of biological database and its integration is a challenge. By effective data visualization techniques, many complex things can be represented and analyzed in an efficient way. The examination of research issues in bioinformatics and the development of new data mining methods for the analysis of bio data has a great significance. Data mining is a powerful technology to find solution for many problems that faced by researchers in their quest to solve problems of our life.

REFERENCES

[1] N Saravanan,&T Devi:A Survey on Biological Databases And Applications of Datamining:Australian Journal of Basic And Applied Sciences, 6(13): 175-180, 2012
[2] Bioinformatics Computing: Bryan Bergeron
[3] Bioinfornatics:R.Sundaralingam&V.Kumaresan
[4] Application of Data Mining in Bioinformatics: Khalid Raza :Indian Journal of Computer Science And Engineering;Vol 1 No 2, 114-118
[5] Advanced Data Mining Technologies in Bioinformatics:Hui-Huang Hsu:Tamkang University, Taipei, Taiwan
[6] Data Mining in Bioinformatics: Study & Surveyof Data Mining And Its Operations in Mining Biological Data: Prof. Sapna V M& Prof. Khushboo Satpute: International Journal of Electronics, Communication & Soft Computing Science And Engineering:Issn: 2277-9477, Volume 2, Issue 9
[7] Http://Pages.Cs.Wisc.Edu/~Oliphant/Cs540/Lecture_Notes/Weka_And_Ilp.Pdf
[8] Introduction to Knowledge Discovery In Databases: Oded Maimon&Lior Rokach:Department of Industrial Engineering,Tel-Aviv University
[9] Integrative Data Mining:The New Direction Inbioinformatics: IEEE Engineering on Medicine And Biology 07395175/01/$10.00©2001ieee
[10] Use of Data Mining in Various Field: A Survey Paper, Smita:,&Priti Sharma:Iosr Journal of Computer Engineering (Iosr-Jce)
[11] Neural Networks -Cathy H Wu, Computer And Information Sciences, Newark, Delaware, USA
Jerry W McCarty, Louisiana State University Health Sciences Center, Shreveport, Louisiana, USA
Li Liao, Computer And Information Sciences, Newark, Delaware, USA; Published Online: September 2010
[12] Neural Networks And Machine Learning in Bioinformatics - Theory And Applications;Esann'2006 Proceedings - European Symposium on Artificial Neural Networks Bruges (Belgium), 26-28 April 2006, D-Side Publi., Isbn 2-930307-06-4
[13] Optimization of Decision Rules in Fuzzy Classification :Renuka Arora& Sudesh Kumar;International Journal of Computer Applications (0975 – 8887) Volume 51– No.3, August 2012
[14] How Can Data Mining Help Biodataanalysis?: Jiawei Han;Department of Computer Scienceuniversity of Illinois At Urbanachampaign
[15] A Survey of Text Mining Techniques And Applications-Vishal Gupta:Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009
[16] Cluster Analysis for Gene Expression Data: A Survey Daxin Jiang Chun Tang&Aidong Zhang
[17] A Survey of Current Work in Biomedical Text Mining :Aaron M. Cohen And William R. Hersh-Henry Stewart Publications 1467-5463. Briefings in Bioinformatics. Vol 6. No 1. 57–71. March 2005
[18] Role of Data Mining in Bioinformatics: A Review:Ashish Shrivastava&Mickey Sahu: Indian Journal of Applied Research