International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 1

Elementary approach towards Biological Data Mining

Faiz Hashmi
Department of Biotechnology, IILM Academy of Higher Learning, Greater Noida, Uttar Pradesh, India

ABSTRACT

In this paper we provide an overview on interactive and integrative knowledge discovery and data mining. The most important challenges, includes the need to develop and apply novel methods, algorithms and tools for the integration, fusion, pre-processing, mapping, analysis and interpretation of complex biomedical data with the aim to identify testable hypotheses, and build realistic models. The HCI-KDD approach, which is a synergistic combination of methodologies and approaches of two areas, Human–Computer Interaction (HCI) and Knowledge Discovery & Data Mining (KDD), offer ideal conditions towards solving these challenges: with the goal of supporting human intelligence with machine intelligence. There is an urgent need for integrative and interactive machine learning solutions, because no medical doctor or biomedical researcher can keep pace today with the increasingly large and complex data sets – often called "Big Data". The application of data mining in the domain of bioinformatics is explained. It also highlights some of the current challenges and opportunities of data mining in bioinformatics.

Keywords: *Data Mining, HCI and KDD (Human-Computer Interaction, Knowledge Discovery & Data Mining), Big Data, Interactive Knowledge discovery.*

Introduction

Data mining is defined as the process of automatically extracting meaningful patterns from usually very large quantities of seemingly unrelated data. Data mining emerges as a new discipline at the end of 1980's. The introduction of new technologies such as computers, satellites, new mass storage media and many others have leads to an exponential growth of collected data. Traditional data analysis techniques often failed to process large amounts of often noisy data efficiently, in an exploratory fashion. The scope of data mining is the knowledge extraction from large data amounts with the help of computers. It is an interdisciplinary areas of researches that has its roots in databases, machine learning and statistics and has contribution from many other areas such as information retrieval, pattern recognition, visualization, parallel and distributed computing. It is an iterative process in which preceding process are modified to support new hypotheses suggested by the data. The main aim of data mining is to explore the databases through automated means and discover meaningful, useful patterns and relationships in data. Data mining can be defined as one particular step of the KDD (knowledge discovery from data) process: the identification of interesting structures in data. It uses different algorithm for classification, regression, clustering or association rules.

The steps for data mining follow the following pattern:

- Data Extraction
- Data Cleansing
- Data Transformation /Reduction
- Data Mining Methods
- Applying Data Mining Algorithm
- Modeling Data
- Pattern Discovery
- Data Visualization

Data Extraction

Data selection and sampling from extracted data by data warehouses, databases data marts repositories is a first challenging step in data mining. Data mining requires a controlled vocabulary, usually implemented as part of a data dictionary, so that a single word can be

used to express a given concept. As millions and thousands of records and variables are gathered in data warehouses and data bases initial mining of meaningful data is quite a complicated process. Typically restrict to computationally enable sample of the holding in an entire data warehouse. The evaluation of the relationships that are revealed in these samples can be used to determine which relationships in the data should be mined further using the complete data warehouse. With large, complex databases, even with sampling, the computational resource requirements associated with non-directed data mining may be excessive. In this situation, researchers generally rely on their knowledge of biology to identify potentially valuable relationships and they limit sampling based on these heuristics.

Data Cleansing

The data collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies.

Once this extracted it has to be preprocessed and cleaned. This is done in following steps:

Data Characterization: It basically deals with documentation of data in an appropriate and meaningful manner, so that any person could understand and interpret the data comfortably. This task is basically done by programmers and other staff involved in data mining project it involves creating a high-level description of the nature and the content of the data to be mined.

Consistency Analysis: It is analyzing the variability of data independent of domain. Based on data values, it is primarily statistical analysis of data. Outliers and values determined to be significantly different from other data may be automatically excluded from the knowledge-discovery process, based on predefined statistical constraints.

For example, data associated with a given parameter that is more than three standard. Deviations from the mean might be excluded from the mining operation.

Domain Analysis: It is validating the data values in a larger context of biology. It is something which goes beyond simply verifying that data value is a text string or an integer, or that it's statistically consistent with other data on the same parameter, to ensure that it makes sense in the context of the biology. Domain analysis requires that someone familiar with the biology create the heuristics that can be applied to the data. Data enrichment: involves strengthening of data from multiple data sources to minimize the limitations of a single data source. It basically involves studying various sources of data For example; two databases on inherited diseases might each be sparsely populated in terms of proteins that are associated with particular diseases. This deficit could be addressed by incorporating data from both databases, assuming only a moderate degree of overlap in the content of the two databases. Frequency and Distribution Analysis: It finds the frequency of occurrence of data during the data mining process by placing the weights on values as a function of their frequency of Occurrence. This is done to maximize the contribution of common findings while minimizing the effect of rare occurrences on the conclusions made from the data-mining output.

Data Transformation

The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc. Normalization: It represents the data in various forms depending on analysis and based on further processes to be implemented. It involves transforming data values from one representation to another, using a predefined range of final values. Various scales are used in normalization process like absolute scales, nominal scales, ordinal scales, rank scales. For example, qualitative values, such as "high" and "low," and qualitative values from multiple sources regarding a particular parameter might be normalized to a numerical score from 1 to 10.

Data Mining

Now we are ready to apply data mining techniques on the data to discover the interesting patterns. The process of data mining is concerned with extracting patterns from the data. Techniques like clustering and association analysis are among the many different techniques used for data mining.

Applying Data Mining Algorithm

This is not a single method or approach, but ii converges various technology and techniques to achieve proper mining of wide range of and also the data of interest biological data. Machine learning methods have wide applicability in data mining algorithms. It includes statistics, biological modeling, adaptive control theory, psychology, and artificial

intelligence (AI).Basically genetic algorithm and neural networks take a major part as a technique to in biological data. Similarly, adaptive control theory, where parameters of System change dynamically to meet the current conditions, and psychological theories, especially those regarding positive and negative reinforcement learning, heavily influence machine learning methods. Artificial Intelligence techniques, such as pattern matching through inductive logic programming, are designed to derive general rules from specific examples.

Data Modeling

Data modeling basically is a process of structuring and organizing the data, and then these structured data are implemented in database management system. Today's biological world demands for heavy exploitation of data These data as are in various forms which has to be capsulated in a meaning full manner .The data are in disparate formats, remotely dispersed, and based on the different vocabularies of Various disciplines. Furthermore, data are often stored or distributed using formats that leave implicit many important features relating to the structure and semantics of the data. Conceptual data modeling involves the development of implementation-independent models that capture and make explicit the principal structural properties of data. Entities such as a biopolymer or a reaction, and their relations, egg catalyzed can be formalized using a conceptual data model. Conceptual models are implementation-independent and can be transformed in systematic ways for implementation using different platforms, e.g. traditional database management systems.

Pattern Discovery

Biology has been transformed from a data poor to a data rich field, with massive accumulation of disparate types of data, for example huge databases of sequences (DNA, RNA, or protein). This data allows important biological insights to be made, partly by finding patterns and motifs that are conserved across many individuals or species; there is now a huge biological literature reporting on such conserved patterns and motifs that have been found in biological datasets. In contrast to the area of pattern matching, the patterns and motifs are generally not known ahead of time, but must be identified or discovered from the data; this task is often very subtle and difficult because the patterns and motifs may be short, may be highly degenerate (containing wildcards and variable length elements), may be ordered differently in different genomes, and are generally hidden in that they make up a small fraction of the data. For particular biological applications, even the definition of a relevant pattern may be difficult to state clearly, or may be unresolved. In bioinformatics, pattern recognition is most often concerned with the automatic classification of character sequences representative of the nucleotide bases or molecular structures, and of 3D protein structures.

Data Visualization

Visualizing biological data is one of the most challenging part of data mining process. In this modern, digital society, how the data is visualized becomes the prime facto, when it comes to communicating or understanding complex concepts. Better the data visualized, better the concepts will be clear. Visualization technologies can provide an intuitive representation of the relationships among large groups of objects or data points that could otherwise be incomprehensible, while providing context and indications of relative importance. The "Sequence Visualization" and "Structure Visualization is types of data visualization techniques.

Houle et al. (2000) refer to a classification of three successive levels for the analysis of biological data, that is identified on the basis of the central dogma of molecular biology:

1. Genomics is the study of an organism's genome and deals with the systematic use of genome information to provide new biological knowledge.

2. Gene expression analysis is the use of quantitative mRNA-level measurements of gene expression (the process by which a gene's coded information is converted into the structural and functional units of a cell) in order to characterize biological processes and elucidate the mechanisms of gene transcription (Houle et al., 2000).

3. Proteomics is the large-scale study of proteins, particularly their structures and functions. These application domains are examined in the following paragraphs. As many genome projects (the endeavors to sequence and map genomes) like the Human Genome Project have been completed, there is a paradigm shift from static structural genomics to dynamic functional genomics (Houle et al., 2000). The term structural genomics refers to the DNA sequence determination and mapping activities, while functional genomics refers to the assignment of functional information to known sequences. There are particular DNA sequences

that have a specific biological role. The identification of such sequences is a problem that concerns bioinformatics scientists. One such sequence is transcription start site, which is the region of DNA where transcription (the process of mRNA production from DNA) starts. Another biologically meaningful sequence is the translation initiation site, which is the site where translation (protein production from mRNA) initiates. Although every cell in an organism -with only few exceptions- has the same set of chromosomes, two cells may have very different properties and functions. This is due to the differences in abundance of proteins. The abundance of a protein is partly determined by the levels of mRNA which in turn are determined by the expression or non-expression of the corresponding gene. A tool for analyzing gene expression is microarray. A microarray experiment measures the relative mRNA levels of typically thousands of genes, providing the ability to compare the expression levels of different biological samples. These samples may correlate with different time points taken during a biological process or with different tissue types such as normal cells and cancer cells (Aas, 2001).

Serial Analysis of Gene Expression (SAGE) is a method that allows the quantitative profiling of a large number of transcripts (Velculescu et al., 1995). A transcript is a sequence of mRNA produced by transcription. However, this method is very expensive in contrast to microarrays, thus there is a limited amount of publicly available SAGE data. One of the concerns of Proteomics is the prediction of protein properties such as active sites, modification sites, localization, stability, globularity, shape, protein domains, secondary structure and interactions (Whishart, 2002). Secondary structure prediction is one of the most important problems in proteomics. The interaction of proteins with other biomolecules is another important issue.

Mining Biological Data

Data mining is the discovery of useful knowledge from databases. It is the main step in the process known as Knowledge Discovery in Databases (KDD) (Fayyad et al., 1996), although the two terms are often used interchangeably. Other steps of the KDD process are the collection, selection, and transformation of the data and the visualization and evaluation of the extracted knowledge. Data mining employs algorithms and techniques from statistics, machine learning, artificial intelligence, databases and data warehousing etc. Some of the most popular tasks are classification, clustering, association and sequence analysis, and regression. Depending on the nature of the data as well as the desired knowledge there is a large number of algorithms for each task. All of these algorithms try to fit a model to the data (Dunham, 2002). Such a model can be either predictive or descriptive. A predictive model makes a prediction about data using known examples, while a descriptive model identifies patterns or relationships in data. Table 3 presents the most common data mining tasks (Dunham, 2002).

Many general data mining systems such as SAS Enterprise Miner, SPSS, S-Plus, IBM Intelligent Miner, Microsoft SQL Server 2000, SGI MineSet, and InxightVizServer can be used for biological data mining. However, some biological datamining tools such as GeneSpring, Spot Fire, VectorNTI, COMPASS, Statistics for Microarray Analysis, and Affymetrix Data Mining Tool have been developed (Han, 2002). Also, a large number of biological data mining tools is provided by National Center for Biotechnology Information and by European Bioinformatics Institute.

Data Mining in Genomics

Many data mining techniques have been proposed to deal with the identification of specific DNA sequences. The most common include neural networks, Bayesian classifiers, decision trees, and Support Vector Machines (SVMs) (Ma & Wang, 1999; Hirsh &Noordewier, 1994; Zien et al., 2000). Sequence recognition algorithms exhibit performance tradeoffs between increasing sensitivity (ability to detect true positives) and decreasing selectivity (ability to exclude false positives) (Houle et al., 2000). However, as Li et al. (2003) state, traditional data mining techniques cannot be directly applied to this type of recognition problems. Thus, there is the need to adapt the existing techniques to this kind of problems. Attempts to overcome this problem have been made using feature generation and feature selection (Zeng& Yap, 2002; Li et al., 2003). Another data mining application in genomic level is the use of clustering algorithms to group structurally related DNA sequences.

Gene Expression Data Mining

The main types of microarray data analysis include (Piatetsky-Shapiro & Tamayo,2003): gene selection, clustering, and classification. Piatetsky-Shapiro and Tamayo (2003) present one great challenge that data mining practitioners have to deal with. Microarray datasets -in contrast with other application domains- contain a small number of records (less than a

hundred), while the number of fields (genes), is typically in thousands. The same case is in SAGE data. This increases the likelihood of finding "false positives". An important issue in data analysis is feature selection. In gene expression analysis the features are the genes. Gene selection is a process of finding the genes most strongly related to a particular class. One benefit provided by this process is the reduction of the foresaid dimensionality of dataset. Moreover, a large number of genes are irrelevant when classification is applied. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied. Clustering is the far most used method in gene expression analysis. Tibshirani et al. (1999) and Aas (2001) provide a classification of clustering methods in two categories: one-way clustering and two-way clustering. Methods of the first category are used to group either genes with similar behavior or samples with similar gene expressions. Two-way clustering methods are used to simultaneously cluster genes and samples. Hierarchical clustering is currently the most frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality. Many techniques such as bootstrap, repeated measurements, mixture model-based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment (Kerr &Churchill, 2001; Yeung et al., 2003; Ghosh&Chinnaiyan, 2002; Smolkin&Ghosh, 2003). In microarray analysis classification is applied to discriminate diseases or to predict outcomes based on gene expression patterns and perhaps even identify the best treatment for given genetic signature (Piatetsky-Shapiro & Tamayo, 2003). Table 4 lists the most commonly used methods in microarray data analysis. Detailed descriptions of these methods can be found in literature (Aas, 2001; Tibshirani et al., 1999; Hastie et al., 2000; Lazzeroni& Owen, 2002; Dudoit et al., 2002; Golub et al., 1999).

Most of the methods used to deal with microarray data analysis can be used for SAGE data analysis. Finally, machine learning and data mining can be applied in order to design microarray experiments except to analyze them (Molla et al., 2004).

Data Mining in Proteomics

Many modification sites can be detected by simply scanning a database that contains known modification sites. However, in some cases, a simple database scan is not effective. The use of neural networks provides better results in these cases. Similar approaches are used for the prediction of active sites. Neural network approaches and nearest neighbor classifiers have been used to deal with protein localization prediction (Whishart, 2002). Neural networks have also been used to predict protein properties such as stability, globularity and shape. Whishart refers to the use of hierarchical clustering algorithms for predicting protein domains. Data mining has been applied for the protein secondary structure prediction. This problem has been studied for over than 30 years and many techniques have been developed (Whishart, 2002). Initially, statistical approaches were adopted to deal with this problem. Later, more accurate techniques based on information theory, Bayes theory, nearest neighbors, and neural networks were developed. Combined methods such as integrated multiple sequence alignments with neural network or nearest neighbor approaches improve prediction accuracy. A density based clustering algorithm (GDBSCAN) is presented by Sander et al. (1998), that can be used to deal with protein interactions. This algorithm is able to cluster point and spatial objects according to both, their spatial and non-spatial attributes.

Databases of Bioinformatics

There are many rapidly growing databases in the field of Bio informatics.

Protein Data Bank

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the PDB.

SWISS-PROT:

SWISS-PROT is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. The SWISS-PROT protein sequence database consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardization purposes the format of SWISS-PROT follows as closely as possible that of the

EMBL Nucleotide Sequence Database. The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria:

(i) Annotations,

(ii) Minimal redundancy and

(iii) Integration with other databases.

Medline

Medical Literature Analysis and Retrieval System Online, or MEDLARS Online is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic Journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution.

The EMBL Nucleotide Sequence Database

The EMBL Database collects, organizes and distributes a database of nucleotide sequence data and related biological information. Since 1982 this work has been done in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Database of Japan (Mishima). Each of the three international collaborating databases DDBJ/EMBL/GenBank, collect a portion of the total sequence data reported world-wide. All new and updated database entries are exchanged between the International Nucleotide Sequence Collaboration on a daily basis. EMBL Database releases are produced quarterly and are distributed on CD-ROM. The most up-to-date data collection is available via Internet and World Wide Web interface.

Biological Data Analysis

Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis:

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Applications of data mining in bioinformatics

- Gene finding,
- Protein function domain detection,
- Function motif detection,
- Protein function inference,
- Disease diagnosis,
- Disease prognosis,
- Disease treatment optimization,
- Protein and gene interaction network,
- Reconstruction,
- Data cleansing,
- Protein sub-cellular location prediction,
- Analysis of protein and nucleotides sequences.

REFERENCES

1. M. Andrade and P. Bork. Automated extraction of Information in molecular biology. FEBS Letters,476:12–17, 2000.
2. T. Attwood and D. Parry-Smith. Introduction to Bioinformatics. Longman Higher Education, 1999.
3. A. Bairoch and R. Apweiler. The SWISS-PROT Protein sequence database and its supplement TrEMBL in 2000. NucleicAcids Res., 28:45–48, 2000.
4. P. Baldi and S. Brunak. Bioinformatics: The MachineLearning Approach, Second Edition. MIT Press, 2001.
5. Bioinformatics Computing By Bryan Bergeron.