International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

Vol.5., Issue.1., 2017
January-February

REVIEW ARTICLE

ISSN: 2321-7758

DATA MINING TOOLS AND APPLICATIONS IN BIOINFORMATICS

SARANGAM KODATI[1], Dr. R VIVEKANANDAM[2]

[1]Research Scholar, Department of Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Science, Sehore,Bhopal,Madhya Pradesh , (India)

[2]Professor, Department of Computer Science and Engineering,Sri Satya Sai University of Technology and Medical Science,Sehore,Bhopal, Madhya Pradesh , (India)

ABSTRACT

There have been a lot about efforts and researches undertaken of developing efficient tools for performing various tasks within data mining. Due to the massive total regarding information embedded of large data warehouses maintained into several domains, the extraction of meaningful pattern is no longer feasible. Data mining mainly contracts with excessive collection on data as inflicts great rigorous computational constraints. In this bill we have focused a different data mining techniques, methods and bioinformatics areas on the research as are helpful or marked as like the important area over data mining Technologies. In that survey a diverse collection on data mining tools are exemplified or also contrasted with the salient features and performance behavior concerning each tool. This paper imparts more number of applications of the data mining and also focuses concerning bioinformatics of the data mining who will useful in the further research.

Keywords*:* Data Mining Techniques, Data Mining Tools, Bioinformatics

1. INTRODUCTION

Data mining is the process of extraction of interesting (nontrivial, implicit, previously unknown and potentially useful) patterns or knowledge from large amount of data. It is the set regarding activities used to find new, hidden or unexpected patterns of data over unusual patterns in data. The integration of biological database is also lacking hence such is very difficult to query more than one database at a time. Therefore, it is important to examine what are the important research issues among Bioinformatics. Data Mining is the exploration and evaluation of large sets, in order to discover meaningful patterns and rules. The key idea is to find effective ways to combine computers power to process data with the human eye's ability to detect patterns. The techniques of data mining are designed for work best with large data sets. Where data relevant to the analysis task are retrieved from the data base ,Data transformation for where data are transformed and consolidated into forms appropriate Knowledge Discovery (KDD) Process – Data mining—core of knowledge discovery process Pattern Evaluation Data Mining Task-relevant Data and Data Warehouse Data Cleaning Data Integration Databases ,Data cleaning for remove Noise and Inconsistent Data ,Data integration for where multiple data sources may be combined ,Data selection for mining by performing summary or aggregation operation ,Data Mining[1]. An essential Process where intelligent methods are applied to extract data patterns, Pattern Evolution to identify the truly interesting patterns representing knowledge based on interestingness measures, Knowledge representation where visualization and knowledge representation techniques are used to present mined knowledge to users

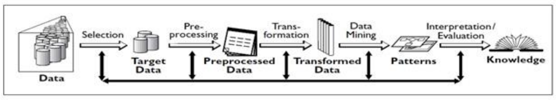

Fig 1: Knowledge Discovery (KDD) Process

2.DATA MINING TOOLS AND APPLICATIONS

The development and application of data mining algorithms requires the use of powerful software tools. As the number over available tools continues to grow, the desire about the most suitable tool becomes increasingly difficult. This paper attempts to support the decision-making method by using discussing the historical development and presenting a range concerning existing state-of-the-art data mining and related tools. Furthermore, we propose criteria for the tool categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, arrival and export options for data and models, platforms, and license policies. These criteria are then used to classify data mining tools into nine different types. The typical characteristics concerning these types are defined and a selection of the most important tools is classifier. Different types of data mining applications are Future Healthcare, Market Basket Analysis, Educational system, Manufacturing Engineering, Customer Relationship and Management, Data analysis techniques for fraud detection, Bioinformatics, Financial Banking, Research Analysis, etc.

Fig 2: Different Types of Data Mining Tools

2.1WEKA

WEKA is the freely available open source [5] application under the GNU. It supports many tasks in data mining such as data preprocessing, classification, clustering and some other process also. The purpose of this application is to utilize the given computer application that allow to perform the machine learning capabilities and useful information is derived to form patterns and trends. It provides an efficient user friendly graphical interface which allows its operation and setup quickly[6]. It is written in C but later the WEKA application has been rewritten into java which can able to survive in almost every computing platform[12].

2.2 Orange

Orange is an open source [8] and also component based software. It is also machine learning software used for an explorative data analysis and visualization, python bindings for scripting. For data preprocessing it provides a set of components that has some exploration techniques and features like modeling, scoring and filtering. It can be implemented in C++ and python.It has the widgets that provide the graphical user's interface for orange's data mining and machine learning methods.

2.3 RATTLE

RATTLE (R Analytical Tool to Learn Easily) is a tool [9] in data mining that gives an uncomplicated and logical interface. It is built on top of the open source and free statistical language R with the help of Gnome graphical interface. This interface takes the user through the basic step of data mining.

2.4jHepWork

jHepWork [13] is an interactive framework for scientific computation, data analysis and visualization which is useful because of scientists, engineers and students. It runs on any operating system where the Java virtual machine can be installed, as the code is written between JAVA. jHepWork is considered amongst five best free and open source data-mining software. It was renamed to SCaVis project from 2013.

2.5 Apache Mahout

Apache Mahout [14] is a project concerning the Apache Software Foundation. The aim of this project is to produce fair implementations of distributed and otherwise scalable machine learning algorithms on the Hadoop Platform. The project mainly focuses regarding collaborative filtering, clustering and classification. Hadoop platforms are beneficial for many implementations

2.6Alpha Miner

Alpha Miner [15] is an open source data mining platform which is designed by the E-Business Technology Institute (ETI). It presents the best cost and performance ratio for data mining applications. Workflow style case construction facilitates simple drag-and-drop operations for general business managers among construction of data mining case. Plugable aspect architecture is very much beneficial into adding new BI capabilities in data import and also because of export, modeling algorithms, model assessment and deployment, data transformation, thus affording extensibility.

2.7 KEEL

KEEL [9] is an open source (GPLv3) java software tool to impose evolutionary algorithms for Data Mining problems. KEEL is designed for solving data mining problems and assessing evolutionary algorithms. The software includes regression, classification, clustering, and sample mining and so on. It permits the user to function a complete analysis on any learning model of comparison to existing ones; that includes a statistical test module for making comparison. It contains a large collection regarding classical advantage of extracting algorithms or preprocessing techniques. Soft-computing methods among knowledge on extracting and learning, and then because of providing scientific and research methods etc.

2.8 Monarch

Monarch is a desktop report mining tool [15] used to extract data beyond human readable report files, such as text, PDF, XPS and HTML. The program was developed by Math Strategies and the software is published by Data watch Corporation 1991. Over 500,000 copies of Monarch have been licensed and the software is between uses in over 40,000 organizations including the latest release version as 11.Monarch can import data from OLE DB/ODBC data sources, spreadsheets and desktop databases.

2.9 TANAGRA

TANAGRA [17] is open source data mining software which is used for purposes like academic and research. It proposes various data mining methods from exploratory analysis of data, statistical and machine learning and for databases area also. The main purpose of Tanagra project is to give researches and students easy-to-use data mining software and allowing analyzing either real or synthetic data.

2.10 SIPINA

SIPINA implements supervised learning [18] and are available for free academic and research purposes. It implements various supervised learning paradigms and especially intended to do decision trees induction. It has some features which have Data Access, Feature Transformation and Selection, Error Evaluation, Classification and Learning Algorithms. SIPINA is mainly a Classification Tree Software. But, other supervised methods are also available such as k-NN, Multilayer perceptions, Naïve Bayes, etc. We can perform the performances comparison and model selection.

2.11 KNIME

KNIME is a perfect fit tool [10] for well-designed which is an effective open source data mining software. It is an open source data analytics, useful for reporting and integrating platform. Using the concept of modular data pipelining it is enable to integrate with various components for machine learning and data mining[11]

2.12 Rapid Miner

Rapid miner [7] is a software platform and open source system for data mining. It has being used in integrated environment which provides facilities for data mining, text mining, machine learning, business analytics and predictive analytics. For integrating the products of data analysis and data mining engine, the rapid miner software is used as a stand-alone application.

3. BIOINFORMATICS AND APPLICATIONS

Bioinformatics is the application of techniques from computer science to problems from biology the term bio medics itself is made up of two parts bio as in biology and informatics as an information. This information can be the genetic code contained in our DNA which is like a recipe book or blueprint for your body telling it how to make everything from muscles to hairs and toes. It can be patient statistics telling us well it's making people sick and what medicines are working.. Or information that allows us to create images of things too small for even the most powerful microscope to see allowing us to study how they behave. In fact sometimes there's so much information that the only way to get through it all is using computers. Interactions of Molecular, Molecular modeling, Analysis of DNA sequence, Analysis of Phylogenetic, Analysis of Protein Sequence ,Drug designing, Molecular Dynamics Simulations ,Tools fields in Bioinformatics[2].

4. CHALLENGES IN BIOINFORMATICS

The implicit goals of bioinformatics are according to read the entire genomes of living things, to identify every gene to match each gene with the protein that encodes and to determine the structure or function of each protein with the help regarding software and techniques. Detailed knowledge about gene sequence, protein structure and function and gene expression pattern to understand how life works at the highest possible resolution. Therefore, Bioinformatics needs to create new and enhanced algorithms because data mining, analysis, comparisons, etc. People with math and programming skills are highly required to bring fresh methods and knowledge. A whole lot on coding should lie done to mangle whole the data but the growth in data along with its increasing complexity has thrown quit a few challenges.

5. DATA MINING APPLICATION USING IN BIOINFORMATICS

Applications of data mining according to bioinformatics include gene finding, protein feature domain detection, characteristic motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein or gene interplay network reconstruction, data cleansing, and protein sub-cellular location prediction. For example, microarray technologies are used according to predict a patient's outcome. On the basis concerning patients' genotypic microarray data, their survival time and risk about tumor metastasis or recurrence can remain estimated[3]. Machine learning be able stay used for peptide identification through mass spectroscopy. Correlation amongst fragment ions among a tandem mass spectrum is crucial amongst reducing stochastic mismatches for peptide identification by using database searching. An efficient scoring algorithm as considers the consistent information among a tunable and comprehensive manners is highly desirable[4].

6. CONCLUSION

In this paper we have discussed the detail study on a number of data mining tasks, techniques, tools, applications and bioinformatics. This review would be useful to researchers to focus regarding the a number of issues concerning data mining. The implementation of data mining strategies will permit users according to retrieve meaningful information from virtually built-in data. These strategies provide variety over applications for industries like retail, telecommunication, Bio-medical etc. These tools predict future traits and behaviors, allowing business to make proactive and present knowledge between the form as is easily understood in accordance with human. The integration over organic databases is also a problem. Challenges of data mining and bioinformatics are fast growing research area today. It is important to examine where are the necessary research issues within bioinformatics and develop data mining techniques for scalable.

7. REFERENCES

[1]. M H Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

[2]. Ramsden, J. (2015). Bioinformatics: An Introduction. 1st ed. Springer.

[3]. Tramontano, A. (2007). Introduction to bioinformatics. 1st ed. London: Chapman & Hall/CRC.

[4]. Zaki , J.; Wang , T.L. and Toivonen, T.T. (2001). BIOKDD01: Workshop on Data Mining in Bioinformatics".

[5]. http://en.wikipedia.org/wiki/Weka_(machine_learning)

[6]. http://www.cs.waikato.ac.nz/ml/weka/

[7]. Rapid Miner [Online]. Available at: http://www.rapidi.com/downloads/tutorial/rapidminer-4.6-tutorial.pdf

[8]. Orange [Online]. Available at: http://www.slideshare.net/jt_4285/data_mining_tool_orange

[9]. Keel [Online]. Available at: http://www.salleurl.edu/GRSI/docs/keel_softcomputing.pdf

[10]. Knime [Online]. Available at: http://www.dataminingresearch.com/index.ph/2010/07/knime_open_sou rce_data_mining_software/.

[11]. Knime [Online]. Available at: http//unipi.it/lib/exe/fetch.php/dm/knime_slides.pdf

[12]. Weka [Online]. Available at: http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[13]. jHepWork [Online]. Available at: http://download.cnet.com/jHepWork/3000_2070 4_75833656.html

[14]. Apache Mahout [Online]. Available at: http://hortonworks.com/hadoop/mahout

[15]. Alpha Miner [Online]. Available at: http://alphaminer.software.informer.com/2.0

[16]. Monarch [Online]. Available at: http://www.ion.icaew.com/ClientFiles/72181ac8-e560- 4bc4-8c20 6be215bef4bc/Monarch%20BI%20Brochure.pdf

[17]. Tanagra [Online]. Available at: http://eric.univ_lyon2.fr/~ricco/tanagra/en.tanagra.html

[18]. Sipina [Online]. Available at: http://eric.univ_lyon2.fr/~ricco/sipina