[COTII- 2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

APPLICATION OF DATA MINING IN ENVIRONMENTAL AND BIOLOGICAL SECTOR

Rajat Verma[*1] **& Dr.Namrata Dhanda**[2]
[*1]M.Tech, Computer Science & Engineering, Amity School of Engineering & Technology, Amity University, Lucknow
[2]Professor, Department of CS/IT, Amity School of Engineering & Technology, Amity University, Lucknow

ABSTRACT

Researchers have used the data mining techniques in various domains. An enormous amount of data has been collected from the scientific domains that include earth sciences, astronomy, meteorology, geology and biological sciences etc. Data mining tools and techniques have been used by researchers in biological and environmental problems also. In biological science, data mining is used in alignment of sequences that is based on the fact that all living organisms are related by evolution. In environmental science data mining is used in prediction of data such as earthquakes and landslide etc. This paper highlights on the research background of protein sequences, (DNA, RNA) sequences, cancer prediction, relational and semantic data mining for biomedical research area. It includes bioinformatics as well as environmental studies.

Keywords: *Bioinformatics tools, data mining tools biological data, environmental data, algorithms, spatial data mining.*

I. INTRODUCTION

Data Mining is basically the extraction of relevant data from a huge amount of data and after this, it should be transformed into an understandable structure so that it can be used further as well. It is a subfield of computer science and has an interdisciplinary prospective. It is the analysis step in the process known as

"Familiarity find in Databases" or "Knowledge Discovery in Databases" (KDD). KDD is a broader concept for finding knowledge in data and focuses on the "bigger level applications" of particular data mining methods.

KDD involves selection, preprocessing, transformation, data mining, interpretation/evaluation and last but not the least knowledge.

Data Mining has a great influence in concern with the biological aspect as well as the environmental aspects.

In Biological aspects, it deals with the protein sequences, multi agent framework, sequences of DNA that is Deoxyribonucleic acid as well as RNA that is known as Ribo nucleic acid. It also works in cancer prediction techniques.

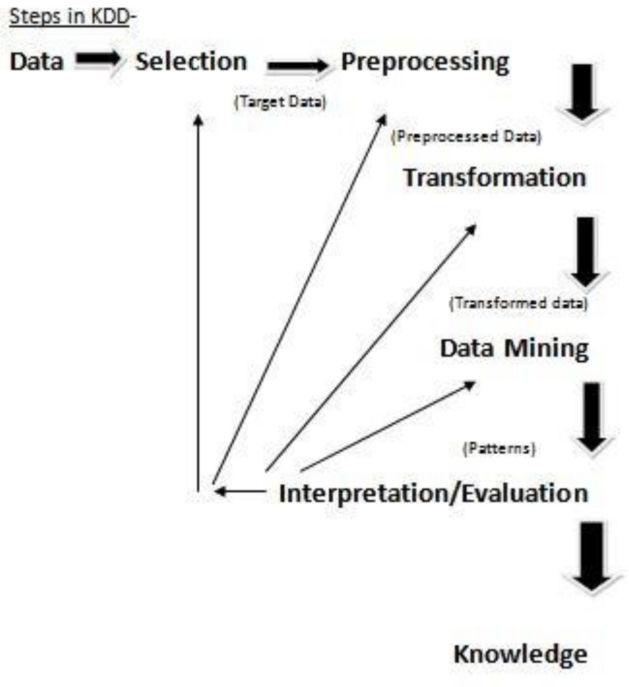

Fig. 1 Knowledge Discovery in Database

In the figure 1 mentioned above the details of the functions that are performed in the section of knowledge discovery in database are as follows-

- Data Cleaning: It is the initial step in which the noise as well as inconsistency is removed from the data. The real world data is full of noise and inconsistency so as the first step of the KDD, cleaning is performed.
- Data Integration: It is the step followed by the cleaning scenario, in which the data from various sources are integrated or combined. Various data sources can be flat files, data base (collection of inter-related records) etc.
- Data Selection: It is the step followed by the integration step in which the data or the unprocessed raw facts or figures are retrieved from the database or the set of inter-related records that are relevant to the task that is being analyzed.
- Data Transformation: It is the step followed by the selection of the data in which the data is converted or transformed into some appropriate forms in relation to the project or objective and the output of this step can directly be sent to the mining scenario. An example of this can be normalized form.
- Data Mining: It is a process in which extraction of relevant data is done from a huge amount of data, particularly considered as a data warehouse or data marts (subset of data warehouse) and the patterns are extracted.
- Pattern Evaluation: It is the step followed by the data mining step and here the evaluation of patterns are done.
- Knowledge Representation: This is basically the last step of the Knowledge Discovery in Data Base and the knowledge is represented.

II. RESEARCH BACKGROUND

A. Brief Review on Biological Sequences in concern to data mining

The researchers have worked on the areas of protein sequences are in [1].
Researchers had developed a hyper clique pattern discovery as the functional modules are highly associated to each other in one or the other way. Hyper clique pattern discovery is an approach to identification of protein complexes. Other researchers who had applied the data mining used the Gen Miner for protein analysis are in [2].

Gen Miner is a preprocessing tool that follows the GIGO principle. In this tool, one can have data from three major protein databases and transform them into a suitable input for Weka and can create the decision tree model. Gen

Miner is an integrated tool that performs preprocessing as well as analysis and provides the following services-

1. Protein behavior discovery
2. Pattern Recognition
3. Integration of multiple tools in 1 program
4. Simple and functional user interface
5. Decision tree

In the area of genome, i.e. the complete set of the genes also the researchers has worked efficiently [3]. It was used for the following-

- DNA Descriptors
- Feature Descriptors
- PCA (Principal Component Analysis)
- Self Organizing Feature Map

For Fuzzy Association Rule mining, i.e. the fuzzy extension in Apriori algorithm, the researchers had worked to a good extent [4].

In concern to the DNA sequences with respect to the data mining prospective many techniques are used such as:

1. DNA data is basically unevenly distributed in nature. For them 2 tools can be used such as:
   - Data Cleaning
   - Data Integration

2. In concern of DNA data analysis, the varied data are put into comparison and similarities are looked for. In this the gene sequences of good or healthy tissues and diseased tissues are put into consideration. This is done by retrieving both tissue and gene sequences and then finding the patterns that are recurring in the entire data.

3. In the field of bio-medical research, the association analysis can be done.

Bioinformatics is new as well as an interesting field in the dimension of science. It is related by science and engineering and the combination of statistics, molecular biology.

Computational models are used to do processing as well as analyzing the biological information of:

- Gene (Physical and functional unit of heredity)
- DNA (Deoxyribonucleic acid)
- RNA (Ribonucleic acid)
- Proteins (Long Chains of amino acid residues)

Role of data mining in other biological segments-
Genetic Algorithms-

Mr. Basheer M Al-Maqaleh and Mr. Hamid Shahbazkia discovered the genetic algorithm [5], for the classification performance to the data that is unknown.
General terms that are involved in this case are-

- Knowledge discovery in databases
- Data Mining
- Machine Learning
- Genetic Algorithm

Researchers who were involved in the Cancer prediction are in [6].

Techniques that were involved in this area-

1. Cancer Prediction Calculator
2. Hidden Markov Model
3. Support Vector Machines
4. Bayesians Networks
5. Association Rules

Researchers who were involved in the Multi-agent framework for Bio-data mining are in [7].
It was based on the framework to demonstrate that how it helps the biologists to do the comprehensive mining task and getting the answers of the biological questions.
For e.g.: With this data mining algorithm it can provide the enquiries of "Leukemia" and "Cancer".

In concern to the (DNA, RNA) sequences, researchers that were involved in this area are in [8].

In these researches multi relational data mining techniques were used.
Non-Negative matrix factorization involved the work of the researchers that are included in [9].It was used for the analysis of biological data.

B. Environmental Problems in concern to the Data Mining Prospective-

An Environment or commonly known as the "Ecosystem" is basically a natural unit of biotic factors as well as abiotic factors. It has influence on the mentioned 3 factors-

- Survival
- Development
- Evolution

There are certain issues of the environmentalists that are related to the natural environment focusing on the climate change, species extinction, pollution and old growth forest loss.

The Environmental issues are given below-

- Earthquakes
- Landslides
- Spatial Data
- Environmental tool

Usually researchers focused on the commercial use but one researcher i.e. K.Muralidharan [10] focused on this particular area.

Actually, his aims were to study the scientific data and he also talked about the earthquake ruptures and volcanoes.
Earthquake prediction section –

3 areas to be looked in this case –

- Ground Water Level
- Chemical changes in ground water
- Radon gas preset in ground water level

The problem of earthquake prediction is based on the data extraction of precursory or prior phenomena.

Researchers that were involved in the risk assessment of Landslides are J-S Lai and T.F Sai.
They took 10 factors into consideration-

- Elevation
- Slope Aspect
- Curvature
- NDVI (Normalized Difference Vegetation Index)
- Fault
- Geology
- Soil
- Land Use
- River
- Road

In concern to Spatial Data Mining, the researchers were involved who proposed algorithms [11].

They used a Database oriented framework for-

- Spatial Data Mining
- Spatial Neighborhood Relations
- Spatial Neighborhood Graphs and their operations
- Spatial Clustering
- Generalized Data Based Clustering
- GDBSCAN
- Applications like Earth Science (5D points) as well as Geography (2-D Polygons) were used.

In the area of geographical knowledge discovery, the researchers that were involved are in [12]
In the areas of text mining, following researchers were involved are in [13].

They worked on the data mining process that is known as "SEMMA".

In SEMMA-

- Sample the data, by using one or more tables. There is no compulsion that we strict the usage of 1 table only.
- Explore the data and find the anticipated relationships.
- Modify the data by creating, selecting and performing other operations.
- Access the data by evaluating the usefulness and reliability of the findings from the process of data mining.

Crisis prediction is also a matter of concern in case of the environmental section that depicts the data intensive process.

The researchers involved are in [14].
Researchers that considered the data mining as a tool to environmental science are depicted in [15].
They used Bayesian Decision Network, Artificial Neural networks.

When it comes to the section of unstructured data environment, the researchers that were involved are in [16].They designed a generalized framework of privacy preservation in the context of distributed data mining.

In the context of classification, regressions as well as the predictive data mining the researchers that were involved are in [17].

III. CONCLUSION

In today's world, an enormous amount of data is collected on a daily basis. So if large amount of data is present then data mining will automatically play an important role. In the same concern, Data mining can also provide tools for the discovery of knowledge from data. As the data mining technique is growing at a fast rate, it can meet the requirements in any area. This paper illustrates about the data mining for environmental and biological problems. This paper highlights on the environmental issues i.e. earthquakes, landslides etc as well as biological issues such as cancer prediction, protein and genomic sequences.

REFERENCES

1. *Xiong, H., Tan, P. N., & Kumar, V. (2006). Hyper clique pattern discovery. Data Mining and Knowledge Discovery, 13(2), 219-242*
2. *Hatzidamianos, G., Diplaris, S., Athanasiadis, I., & Mitkas, P. A. (2003, November). GenMiner: A data mining tool for protein analysis. In Proceedings of the 9th Panhellenic Conference on Informatics, Thessaloniki, Greece.*
3. *Sen, S., Narasimhan, S., & Konar, A. (2007). Biological Data Mining for Genomic Clustering Using Unsupervised Neural Learning. Engineering Letters, 14(2), 61-71.*
4. *Soni, S., & Vyas, O. P. (2012). Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(1), 11-22.*
5. *Al-Maqaleh, B. M., & Shahbazkia, H. (2012). A genetic algorithm for discovering classification rules in data mining. International Journal of Computer Applications, 41(18).*
6. *Mokharrak, W., Al Khalaf, N., & Altman, T. (2012, January). Application of Bioinformatics and Data Mining in Cancer Prediction. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).*
7. *Yang, P., Tao, L., Xu, L., & Zhang, Z. (2009, July). Multiagent Framework for Bio-data Mining. In RSKT (pp. 200-207).*
8. *Page, D., & Craven, M. (2003). Biological applications of multi-relational data mining. ACM SIGKDD Explorations Newsletter, 5(1), 69-79.*
9. *Li, Y., Rezaei, B., Ngom, A., & Rueda, L. (2015, August). Prediction of high-throughput protein-protein interactions based on protein sequence information. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on (pp. 1-6). IEEE.*
10. *Shrivastava, P., & Shukla, M. (2013). A brief survey on data mining for biological and environmental problems. International Journal of Scientific & Engineering Research, 4(11).*
11. *Ester, M., Kriegel, H. P., & Sander, J. (2001). Algorithms and applications for spatial data mining. Geographic Data Mining and Knowledge Discovery, 5(6).*
12. *Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. Computers, Environment and Urban Systems, 33(6), 403-408.*
13. *Dave Smith, S. A. S., & Marlow, U. K. (2007). Data Mining in the Clinical Research Environment.*
14. *Krammer, P., Šeleng, M., Habala, O., & Hluchý, L. Advanced Data Mining Methods for Environmental Crisis Prediction.*
15. *Spate, J. M., Gibert, K., Sànchez-Marrè, M., Frank, E., Comas, J., & Athanasiadis, I. N. (2006). Data Mining as a tool for environmental scientists.*
16. *Thavavel, V., & Sivakumar, S. (2012). A generalized framework of privacy preservation in distributed data mining for unstructured data environment. IJCSI International Journal of Computer Science Issues, 9(1), 1694-0814.*
17. *Velickov, S., & Solomatine, D. (2000, March). Predictive data mining: practical examples. In 2nd Joint Workshop on Applied AI in Civil Engineerin.g*