A Brief Survey On Data mining For Biological and Environmental Problems

Pooja Shrivastava Dr.Manoj Shukla

Abstract-In the past, many researchers used data mining techniques in any area. A lot of amounts of data have been collected from scientific domains such as geo sciences, astronomy, meteorology, geology and biological sciences. Data mining techniques and tools used by researchers in biological and environmental problems also. In biological science data mining used in sequences alignment is based on the fact that all living organisms are related by evolution and in environmental science data mining used in predicting data such as earthquakes and landslide etc. This paper highlights on the wide survey of protein sequences, (DNA, RNA) sequences, cancer prediction, relational and semantic data mining for biomedical research area. Health care data, multiagent framework for bio data mining, predicting earthquakes, landslide and spatial data in distributed data mining algorithms and tools. This is including in bioinformatics and environmental studies also.

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

1-Introduction

In my survey researcher discussed protein sequences which protein who highly affiliated with each other and discovered a Hyperclique pattern approach for extracting functional modules. Gen Miner is a pre-processing software tool and that can receive data from protein data base and transform them in a form suitable for input WEKA software. Decision tree model and WEKA tool used in this protein analysis researchers apply some classification technique like classification technique like neural networks, genetic algorithm Fuzzy ARTMAP, Rough set classifier. It is the new technologies such as computers. Protein analysis together these keywords are searched on genome sequences also. In genome sequences focused on DNA-descriptors, feature descriptors principal component analysis (PCA), and self-organizing feature Map (SOFM),genomic data mining is an important problem in bioinformatics. Introduced DNA-descriptor for a sample of the DNA sequence of mouse in the bar-diagram form. Fuzzy association rule mining and fuzzy weighted associative classifier (FWAC) used in a predictive technique for health care data mining. In this related work used Association rule mining, weighted association rule mining etc. Genetic algorithm used for classification rules in data mining. Data mining researcher focused on the commercial sectors and application only few researcher works in this particular area. Environment problem is also big deal for data mining. But usually in this area data mining used such as predictive data like land slide, earthquakes, spatial data etc. So many tools, algorithms, concepts and techniques used of data mining. We can see that in any companies and any other government and private sectors data traffic problem is created so researcher given the environmental tool also to solve these problems. Many type of keyword introduce in this survey paper.

Pooja Shrivastava
Computer Science and Engineering
Jayoti Vidyapeeth Women's University
Ph.D. Research Scholar (JVR-II/12/5010) Jaipur, India
Srivastavapooja547@gmail.com

Dr. Manoj Shukla
Computer Science and Engineering
Sunder Deep Group of Institution
Associate Professor, Ghaziabad U.P. India
mkshukla001@gmail.com

- Bioinformatics
- Earthquakes
- Land slide
- Spatial data
- Biological sequence
- Cancer prediction

In above these key word we distributed the data mining any type of area and improve the concepts and techniques of data mining. In earthquakes problem worked hierarchical and non hierarchical clustering method. Risk assessment and validation used for land slide and found the accuracy. In biological sequence is so difficult for analyzing the protein and genomic sequences. But researcher accepted these challenge and reduced these sequence problem something. Data mining used other biological terms also and these concepts highlighted any other term of bioinformatics and environment. It has used the spatial information technologies, hazard assessments, monitor, decision making support, mitigation, and spatial data. In below we can see that researcher given the spatial data mining structure. These structure shown the how the spatial data used in data mining and created the knowledge discovery from data base. First of all user face, mine, data resource. Now we going on review of letter in these problems.

2- A Brief review of latter data mining for Biological and Environmental problems

2.1 Survey on biological sequences in data mining:

Hui xiong, Xiaofeng HE ,Chris ding ,Ya Zhang, Vipin kumar, Stephen r.holbrook[1] worked on identifications of functional modules in protein complexes. highly affiliated to each other and do not act isolated in cell but with function proteins work in cellular path ways in pair as a component. So researcher developed hyperclique pattern discovery from protein complexes it is type of association patterns.

A) Hyper clique pattern discovery is based on association rule.
B) Protein complex data and analysis tools(Gene ontology)
C) Analysis of hyperclique pattern using Gene Ontology
D) Hyper clique pattern as functional modules

Proteins are

So researchers have been developed in these pattern for protein complexes.

In this research described hyper clique pattern discovery approach to identification of proteins complexes. Any other researcher applied data mining tool for analysis the protein **Gerasimos Hatzidamianos, Sotiris Diplaris, Ioannis Athanasiadis, Pericles A. Mitkas [2]** have been dealt with Gen miner. It is a pre-processing software tool. with the help of this tool we can receive data from three major protein databases and transform them in a form suitable for input to the WEKA data mining suite and created the decision tree model with used the derived training set and efficiency test was conducted in this research. In this research tried to solve problem "Given an amino acid database or training set that exists in proteins with known properties (that have been experimentally specified), we aim to create a tool that can classify new, unknown proteins in some known to the training set family of proteins, referred as protein class." Researcher was used weka tool.with help of weka tool protein data sets and SQL manager also used in this research.

From a functional point of view, GenMiner offered various services that are presented below:

1. Protein behavior discovering
2. Protein recognition
3. Decision tree building
4. Simple and functional user interface
5. Integration of multiple tools in one program

Researcher have been designed **a** scheme for automatic identification of a species from its genome sequence. For a DNA sequences many technologies was used of data mining. **Shreyas Sen, Seetharam Narasimhan, and Amit Konar [3]** have been produced the Clustering Using Unsupervised Neural Learning for genomic sequence. It was used DNA-descriptors, Feature Descriptors, Principal Component Analysis (PCA), Self-Organizing Feature Map (SOFM). That was very challenging task for the researcher. But with the help of used the biological data mining it was implement in easy way. Bioinformatics is a new field of science. It is related by the science and engineering. And the combination of statistics, molecular biology, and computational methods is used for analyzing and processing biological information like gene, DNA, RNA, and proteins and extract the other information.

2.2 Concepts and technologies of data mining used in others Biological segments:

Researchers are used data mining techniques uses other biological factors. This phenomena is popular and useful. **Sunita Soni and O.P.Vyas[4]** have been dealt with problem of classification using Fuzzy Association

Rule Mining. And these proposed the concept of Fuzzy Weighted Associative Classifier (FWAC). This research based on A PREDICTIVE TECHNIQUE FOR HEALTH CARE DATA MINING. In this research was used keywords Associative Classifiers, Fuzzy Weighted Association Rule, FWAC, Fuzzy weighted support, Fuzzy weighted Confidence. In this research related work is Association Rule Mining, Weighted Association Rule Mining, Fuzzy Association Rule Mining (FARM), Fuzzy Weighted Association Rule Mining, Incorporating Weight in ARM, Utilizing Weight in Medical Domain, Fuzziness of Quantitative Attribute. Two important modification have been proposed (weight of an attribute and fuzzy fication of quantitative attributes). And these problem refused by Fuzzy Attribute Weight (FAW), Fuzzy Attribute set Transaction Weight (FASTW), Fuzzy Attribute Set Weight (FASW) Fuzzy Weighted Support (FWS) and Fuzzy Weighted Confidence (FWC) for Fuzzy Weighted Associative Classifiers (FWAC). Technique for Fuzzy Weighted Association Rule Mining is known as (FWARM). **Basheer M. Al-Maqaleh and Hamid Shahbazkia[5]** have been discovered Genetic algorithm for classification performance to unknown data. It have general terms is Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning, Genetic Algorithm. And used the techniques are Classification Rule, Genetic Operators, Fitness Function, Predictive Accuracy. This work proved that none of the selected learners improved the predictive accuracy on any dataset, as much as the proposed algorithm did. Data mining used in other biological data sets. **Wafa Mokharrak, Nedhal Al Khalaf, Tom Altman[6]** Cancer prediction techniques are very difficult many concepts used in research of data mining. Such as decision tree, Cancer prediction calculator, Hidden Markov Model, Support vector machine, Network approach for cancer, Bayesian Networks, Association rules. In this research important role of
Bioinformatics and data mining. **Pengyi Yang, Li Tao, Liang Xu, and Zili Zhang [7]** proposed the Multiagent framework for Bio-data mining. It is based on the framework, researcher developed a prototype system to demonstrate how it helps the biologists to perform a comprehensive mining task for answering biological questions. It is new research and understood to easy. With the help of data mining algorithm it provided the enquiries of "Leukemia" and "Cancer". This is multiagent based on bio-data mining framework will help in future to bridge the knowledge gap between data mining community and biology community, and enhance the reusability of biological databases as well as data mining algorithms. **David Page and Mark Craven [8]** worked on biological applications of multi relational data mining first of all biological data base contain a many type of variety of data sets and this data sets relation is found. So multi relational data mining techniques applied in this research. In protein sequence and (DNA,RNA) sequences used this techniques. Multi relational algorithms are applied protein and genomic sequences and developed the new research direction of multi relational data mining algorithms. **Yifeng Li and Alioune Ngom [9]** introduced the Non-negative matrix Factorization for logical data mining. NMF important approach to analyze the biological data. In this research implemented R and programming language, and MATLAB toolbox. With the help of this toolbox, data mining approaches such as techniques clustering, biclustering, feature extraction, feature selection, classification, overcoming missing values, visualization, and statistical comparison can be easily done. With the help of this research we can improve this toolbox and include more NFM algorithm such as NMF LS-NMF and implement statistical comparison methods in future.

2.3 include the environment problems also in data mining:

First of all what is environment **environment** is the biotic and abiotic surrounding of an organism, or population, and includes particularly the factors that have an influence in their survival, development and evolution. Environmentalism is a broad social and philosophical movement that, in a large part, seeks to minimise and compensate the negative effect of human activity on the biophysical environment. The issues of concern for environmentalists usually relate to the natural environment with the more important ones being climate change, species extinction, pollution, and old growth forest loss. Skilled environment scientists have an important role to play in examining various environmental problems in a scientific manner and carry out R&D activites for developing cleaner techonologies and promoting sustainable development. The Enviroment problem given in below:

- Earth quakes
- Land slide

- Spatial Data
- Environmental tool

In above we can see that any researcher have been focused on environmental problem and improved the solution for reduced the problem. Data mining researcher focused on the only commercial data. Only few researchers focused on this field. **K. Muralidharan [10]** worked on this particular area. His introduced predicting earth quakes using data mining. His aims study the scientific data. This research highlighted the data mining techniques applied to mine for surface changes over time (e.g. Earthquake rupture) and with the help of data mining change in the intensity of volcans. Statistical model used in this area and highlighted observable space time earthquake patterns from unobservable dynamics using data mining techniques, pattern recognition and ensemble forecasting and given how data mining finding the why predict the earth quake.

2.3.1 EARTHQUAKE PREDICTION:

- Ground water levels
- Chemical changes in Ground water
- Radon Gas in Ground water wells.

Hierarchical and non hierarchical method used include in this research. The problem of earthquake prediction is based on data extraction of precursory phenomena and it is very challenging task but data mining tools and method used.

J-S. Lai , F.Tsai [11] verification and risk assessment for land slide in the Shimen Reservoir watershed of Taiwan using spatial data analysis and data mining In this study, eleven factors were considered, including elevation (Digital Elevation Model, DEM), slope, aspect, curvature, NDVI (Normalized Difference Vegetation Index), fault, geology, soil, land use, river and road and find the result accuracy and kappa coefficient. This paper presented decision tree algorithm, spatial data, mechanism for filtering uncertain data from the data sets to improve the reliability of landslide predictions and given the accuracy and kappa coefficient in verification can reach 98.1 % and 0.8829. and given the more reliable result in the study site. In our life we are focused on the spatial data such as scientific data. Knowledge discovery in databases (KDD) has been defined as the non-trivial process of discovering valid, novel, potentially useful and ultimately understandable patterns from data.

Martin Ester, Hans-Peter Kriegel, Jorg Sander (University of Munich)[12] have been proposed algorithms and application for spatial data mining. In this researcher have introduced A Database-Oriented Framework for Spatial Data Mining, Spatial Neighbourhood Relations, Spatial Neighbourhood Graphs and their Operations, Spatial Clustering, Generalized Density-Based Clustering, Algorithm GDBSCAN and many applications has used.

Application 1: Earth Science (5D points)

Application 2: Geography (2D polygons)

In this research was introduction database-oriented framework for spatial data mining which is based on the concepts of neighbourhood graphs and paths and many applications of DBMS and data mining algorithms and concepts. **Diansheng Guo , Jeremy Mennis [13]** worked on this area. This research focused on geographical data and spatial data by developing new techniques for point pattern analysis, prediction in space–time data, and analysis of moving object data, as well as by demonstrating applications of genetic algorithms for optimization in the context of image classification and spatial interpolation. This is based on the new technique and in future we can improve that. Data mining has wide adoption in many area and it is so popular in many industry. **Dave Smith, SAS, Marlow, UK [14]** discussed the topic of data and text mining in general, before focusing on applications in the clinical research field. His introduced the data mining process SEMMA.

- **Sample** the data by creating one or more data tables.
- **Explore** the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
- **Modify** the data by creating, selecting, and transforming the variables to focus the model selection process.
- **Modify** the data by creating, selecting, and transforming the variables to focus the model selection process.
- **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.

Seema is not complete data mining it is miner tools and caryying out the data mining. Semma is not data mining methodology. SAS has developed its own methodology. Data mining and text mining are powerful techniques and we can add understanding to many facotrs of the clinical research environment.

Peter Krammer, Martin seleng, Ondrej Habala, and Ladislav Hluchy [15] have been focused on the crisis prediction means how the data-intensive processes can be applied to benefit the experts and any other thing such as daily whether prediction and time critical assessment for environmentally events and described data integration, data pre-processing, model training process and used architecture in detail. Research dealt with architecture and data mining on the prediction of the floods on the Orava basin and river and also dealt with deals with numerical prediction data mining in hydrology area. I future we want to improve flood experiment as well as crisis prediction**. Jessica Spate, Karina Gibert, Miquel S`anchez-Marre, Eibe Frank, Joaquim Comas, Ioannis Athanasiadis , Rebecca Letcher [16]** given the data mining as a tool for environmental science. We can see that data traffic problem is so rich and it is very harmful problem in industry so researcher used the data mining concepts and technique as a tool. His was propose the Certain techniques such as Artificial Neural Networks, Clustering, Case-Based Reasoning and more recently Bayesian Decision Networks have found application in environmental modelling while other methods.

V.THAVAVEL and S.SIVAKUMAR [17] worked on the unstructured data environment. His have designed generalized Framework of Privacy Preservation in distributed data mining. Unstructured data is unsolved the problems in information industry and data mining paradigm. This proposed a solution to this problem by managing unstructured data in to structured data using legacy system and distributed data partitioned method for gives distributed data for mining multi text documents and frame work gives the testing of the similarities among text documents and privacy preserving meta data hiding technique, which are explored in text mining. This research provided the frame work for privacy preservation of Meta data using hiding technique in unstructured data environment with a distributed mechanism. Predictive data mining is most important part of the data mining area. **Slavco Velickov and Dimitri Solomatine [18]** given the data mining piratical example in own research. In this paper introduced the problem of classification and regression and solve these problem increase the dimensions and complexity and highlighted the methodologies for predictive data mining. Bayesian classifier, decision tree induction algorithm (C4.5) and 'local' modelling using chaos theory in this research and Last part of the paper presented application of the predictive data mining techniques to hydro-meteorological data.

3- Conclusion

We live in a world where vast amounts of data are collected daily. such data is an important to need so data mining play the important role. Data mining can meet this need by providing tools to discovery knowledge from data. Now days we can see that data mining use in any area. This paper presents survey on the data mining for biological and environment problem. So we observe that many kind of concepts and technique used in these problems. And try the removed complicated and hard type of data. This paper highlights on biological sequences problem such as protein and genomic sequences and other biological segments such as cancer prediction. In environment presents earth quakes, land slide, spatial data and environmental tool also discuss. Data mining algorithms, tools and concepts used in these problems Such as MATLAB, WEKA, SWIISPORT , Clustering , Biclstering and any other thing in this survey.

References

1- Hui xiong, Xiaofeng HE ,Chris ding ,Ya Zhang, Vipin kumar, Stephen r.holbrook.

2- Gerasimos Hatzidamianos, Sotiris Diplaris, Ioannis Athanasiadis, Pericles A. Mitkas.

3- Shreyas Sen, Seetharam Narasimhan, and Amit Konar[Engineering Letters, 14:2, EL_14_2_8 (Advance online publication: 16 May 2007].

4- Sunita Soni and O.P.Vyas[International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012].

5-Basheer M. Al-Maqaleh *and* Hamid Shahbazkia [International Journal of Computer Applications (0975 – 8887) Volume 41– No.18, March 2012].

6- Wafa Mokharrak, Nedhal Al Khalaf, Tom Altman[Department of Computer Science and Engineering, University of Colorado Denver, Denver, Colorado, United States of America].

7-Pengyi Yang, Li Tao, Liang Xu, and Zili Zhang[P. Wen et al. (Eds.): RSKT 2009, LNCS 5589, pp. 200–207, 2009. _c Springer-Verlag Berlin Heidelberg 2009].

8- David Page and Mark Craven[Dept. of Biostatistics and Medical Informatics and Dept. of Computer Sciences].

9-Yifeng Li and Alioune Ngom[School of Computer Science, University of Windsor, Windsor, Ontario, Canada].

10- K.Muralidharan[II Year B.Tech Information Technology Karpagam Institute of Technology COIMBATORE – 21].

11- J-S. Lai , F.Tsai[International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B2, 2012 XXII ISPRS Congress, 25 August –01 September 2012, Melbourne, Australia].

12- Martin Ester, Hans-Peter Kriegel, Jorg Sander (University of Munich)[Published in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, Taylor and Francis, 2001].

13- Diansheng Guo , Jeremy Mennis[Computers, Environment and Urban Systems 33 (2009) 403–408].

14- Dave Smith, SAS, Marlow, UK [Data Mining in the Clinical Research Environment [push 2007].

15- Peter Krammer, Martin Šeleng1, Ondrej Habala1, and Ladislav Hluchý[Accessed Feb.2012].

16- Jessica Spate, Karina Gibert, Miquel S`anchez-Marre, Eibe Frank, Joaquim Comas, Ioannis Athanasiadis , Rebecca Letcher.

17- V.THAVAVEL and S.SIVAKUMAR[IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012 ISSN (Online): 1694-0814 www.IJCSI.org].

18- Velickov and Dimitri Solomatine[Artificial Intelligence in Civil Engineering. Proc. 2nd Joint Workshop, March 2000, Cottbus, Germany. ISBN 3-934934-00-5].