Application based technical Approaches of data mining in Pharmaceuticals, and Research approaches in biomedical and Bioinformatics

[1]Sayyada Sara Banu, , [2]Dr.Perumal Uma, [3]Mohammed Waseem Ashfaque , [4]Quadri S.S Ali Ahmed
[1] . College of computer science and information system, Jazan university, Saudi Arabia
email:- sayyada.sara@gmail.com
[2].College of computer science and information system, Jazan university,Saudi Arabia.
email: prmluma@gmail.com
[3]. Department of Computer Science & IT, College of Management and Computer Technology, Aurangabad, India
email: waseem2000in@gmail.com
[4].Department of Computer Science & IT, College of Management and Computer Technology, Aurangabad, India.
email:- aliahmedquadri@yahoo.co.in

Abstract-In the past study shows that flow of direction in the field of pharmaceutical was quit slow and simplest and by the time the process of transformation of information was so complex and the it was out of the reach to the technology, new modern technology could not reach to catch the pharmaceutical field. Then the later on technology becomes the compulsorily part of business and its contributed into business progress and developments. But now a days its get technology enabled and smoothly and easily pharma industries managing their billings and inventories and developing new products and services and now its easy to maintain and merging the drugs detail like its cost ,and usage with the patients records prescribe by the doctors in the hospitals .and data collection methods have improved data manipulation techniques are yet to keep pace with them data mining called and refer with the specific term as pattern analysis on large data sets used like clustering, segmentation and classification for helping better manipulation of the data and hence it helps to the pharma firms and industries this paper describes the vital role of data Mining in the pharma industry and thus data mining improves the quality of decision making services in pharmaceutical fields. This paper also describe a brief overviews of tool kits of Data mining and its various Applications in the field of Biomedical research in terms of relational approaches of data minings with the Emphasis on propositionalisation and relational subgroup discovery, and which is quit helpful to prove to be effective for data analysis in biomedical and its applications and in Bioinformatics as well.

Keywords— *Data Mining; drug discovery; pharmacy industry; relational data mining; semantic data mining; biomedicine.*

*****

I. INTRODUCTION

1.1 Definition and background

Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems[1] Traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive and highly subjective[2] Data Mining, popularly called as knowledge discovery in large data, enables firms and organizations to make calculated decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools, Data mining the life sciences researcher to mine data to understand safety and efficacy profiles within the patient population. By tackling the question of patient selection within the framework of demonstrating groups that are most responsive, Data mining is sure to penetrate the drug development marketplace. Data mining framework enables specialists to create customized nodes that can be shared throughout the organization, making the application attractive to skilled modelers in a pharmaceutical company's bioinformatics division. The paper discusses how Data Mining discovers and extracts useful patterns from this large data to find observable patterns. The paper demonstrates the ability of Data Mining in improving the quality of decision making process in pharma industry. Data analysis in biomedical applications aims at extracting potentially new relationships from data and providing Insightful representations of detected relationships. Methods for symbolic data analysis are preferred since highly accurate but non-interpretable classifiers are frequently considered useless for medical practice. Subgroup discovery techniques [3, 4] are of interest to biomedical research, as they enable the discovery of patient subgroups from classified patient data,Bioinformatics, is a field committed to the interpretation and analysis of biological data using computational techniques, has evolved tremendously in recent years due to the explosive growth of biological information generated by the scientific community. Bioinformatics is the science of managing, mining, integrating, and interpreting information from biological data at the genomic, proteomic, phylogenetic, cellular, or whole organism levels. The need for bioinformatics tools and expertise has increased as genome sequencing projects have resulted in an exponential growth in complete and partial sequence databases. Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. [5]. The present article provides an overview of

the available literature on data mining, and its aspects. This is followed by Section -2 which discusses the state of art of soft computing and we discuss each of the soft computing methods in brief. finally we conclude the significance of soft computing in data mining is highlighted.[5][6].The purpose of this paper is to provide an overall understanding of Data mining and soft computing techniques and their application and usage in bioinformatics

Fig:1 Data mining various process

1.2 Data Mining Techniques.

Pharma industries rely on decision oriented, systemic selection models that enable the decision maker to evaluate the payoff that is expected to result from the implementation of a proposed selection program. Such models go beyond an examination of the size of the validity coefficient and take a host of issues such as capital budgeting and strategic outcomes at the group and organizational levels. Many organizations generate mountains of data about their new drugs discovered and its performance reports, etc. This data is a strategic resource. Now, making use of most of these strategic resources will lead to improving the quality of pharma industries. give six important steps in the Data Mining process as

1. Problem Definition.
2. Knowledge acquisition.
3. Data selection.
4. Data Preprocessing.
5. Analysis and Interpretation.
6. Reporting and Use.

Identify the Data Mining process as

1. Definition of the objectives of the analysis.
2. Selection &Pretreatment of the data.
4. Explanatory analysis.
5. Specification of the statistical methods.
6. Analysis of the data.
7. Evaluation and comparison of methods.
8. Interpretation of the chosen model.

The techniques and methods in Data Mining need brief mention to have better understanding. [7]

1.3 Relational data mining for biomedical applications

We first present selected approaches to inductive logic programming (ILP) [8, 9] and relational data mining (RDM) [10] which showed a great potential for biomedical research due to their capacity of using background knowledge in the learning process. From the available background knowledge (encoded as logical facts or rules) and a set of classified examples (encoded as a set of logical facts), an ILP/RDM algorithm derives a hypothesized logic program which explains the positive examples. While ILP focuses on data and background knowledge represented in a logical formalism, RDM assumes that the background knowledge and data are encoded in a unique relational database format. Compared to standard data mining techniques where the input data is typically stored in a single data table (e.g., in Excel), the input to an ILP/RDM algorithm is thus much more complex. Propositionalization [11] is a RDM approach, which has been applied in several biomedical applications. Consider relational subgroup discovery, an approach effectively implemented in the RSD algorithm [12]. RSD generates descriptive rules as conjunctions of terms which encode background knowledge concepts. RSD performs example-weighting [13] (used in the so-called weighted covering algorithm) and uses the weighted relative accuracy (WRAcc) measure as a heuristic for rule selection. For example, an induced description of gene group A, discovered by RSD for the CNS (central nervous system) cancer class in the problem of distinguishing between 14 cancer types determines group A of differentially expressed genes in CNS as a conjunction of two relational features [14]: general Group(A) fi(A)&fk(A), where the two features, fi(A) and fk(A), constructed in the

II. LITERTURE REVIEW

2.1 Clustering.

It is a method by which similar records are grouped together. Clustering is usually used to mean segmentation. An organization can take the hierarchy of classes that group similar events. Using clustering, employees can be grouped based on income, age, occupation, housing etc. In business, clustering helps identify groups of similarities; characterize customer groups based on purchasing patterns, etc.

2.2 Data Mining and Statistics.

The ability to build a successful predictive model depends on past data. Data Mining is designed to learn from past success and failures and will be able to predict what will happen next (future prediction). One may think why use Data Mining in pharma industry organizations when statistical analysis is already been performed. The Data Mining tool checks the statistical significance of the predicted patterns and reports. Data Mining will tell that it is likely that something unlikely [15] will happen. If Data Mining tool finds that 100 percent of the drugs of some particular large group have included for the performance analysis, but among them only 10 drugs have the characteristics of high performance ratings, then the tool can warn that it is very likely to be an idiosyncrasy of the data base rather than a usual predictive pattern

2.3 Applications Of Data Mining in the Pharmaceutical Industry

Most healthcare institutions lack the appropriate information systems to produce reliable reports with respect to other information than purely financial and volume related statements [16].The management of pharma industry starts to recognize the relevance of the definition of drugs and products in relation to management information. In the turmoil between costs, care-results and patient satisfaction the right balance is needed and can be found in upcoming information and Communication technology. The delivery of healthcare has always been information intensive, and there are signs that the industry is recognizing the increasing importance of information processing in the new managed care environment [17]. Most automated systems are used as a tool for daily work: they are focused on 'production' (daily registration). All the data, which are used to keep the organization running, operational data, are in these automated systems. These systems are also called legacy systems. There is a growing need to do more with the data of an organization than to use them for administration only. A lot of information is hidden in the legacy systems. This information can easily be extracted. Most of the times this cannot be done directly from the legacy systems, because these are not build to answer questions that are unpredictable. Research shows that [18] [19] that successful decision systems enriched with analytical solutions are necessary for healthcare information systems Given the size of the databases being queried, there is likely to be a trade-off in accuracy of information and processing time. Sampling techniques and tests of significance may be satisfactory to identify some of the more common relationships; however, uncommon relationships may require substantial search time.The amount of existing pharmaceutical information (pharmacological properties, dosages, contraindications, warnings, etc.) is enormous; however, this fact reflects the number of medicines on the market, rather than an abundance of detailed information about each product. Data mining techniques have been used in astronomy, bioinformatics, drug discovery and many more Many organizations now employ data mining as a secreat weapon.

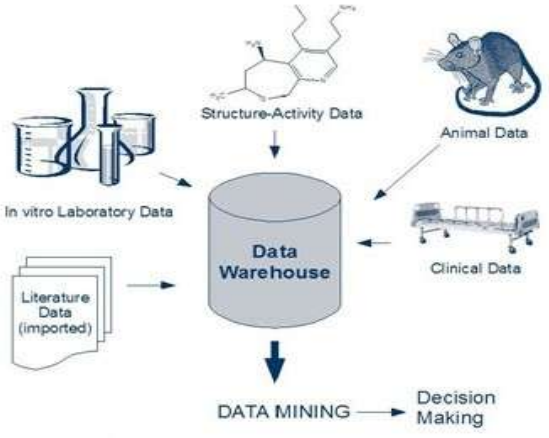

Fig:2 Data mining in various applications

2.3.1 Various Science applications

.

2.3.2 Business applications

to keep in pace or gain a competitive edge .Data mining has been used in advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care, etc.

2.3.3 Other application

Data mining has been successfully used for various
other application such as Web: search engines,, Government ,law enforcement, profiling tax cheaters,
anti-terror, credit approval, etc. Putting it together Data Mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/evaluation step of the KDD process

2.4 *Development of new drugs.*

This research need to use data mining tools and techniques. This can be achieved by clustering the molecules into groups according to the chemical properties of the molecules via cluster analysis [20]. This way every time a new molecule is discovered it can be grouped with other chemically similar molecules. This would help the researchers in finding out with therapeutic group the new molecule would belong to. Mining can help us to measure the chemical activity of the molecule on specific disease say tuberculosis and find out which part of the molecule is causing the action. This way we can combine a vast number of molecules forming a super molecule with only the specific part of the molecule which is responsible for the action and inhibiting the other parts. This would greatly reduce the adverse effects associated with drug actions

2.5 Development tests and predicts Drug behavior

There many issues which affect the success of a drug which has been marketed which can impact the future development of the drug. Firstly adverse reactions to the drugs are reported spontaneously and not in any organized manner. Secondly we can only compare the adverse reactions with the drugs of our own company and not with other drugs from competing firms. And thirdly we only have information on the patient taking the drug not the adverse reaction that the patient is suffering from. All this can be solved with creation of a data warehouse for drug reactions and running business intelligence tools on them a basic classification tool can solve much of the problems faced here. We could find out the adverse reactions associated with a specific drug and still go a step further to show if any specific condition aggravates the adverse reaction for eg age, sex, and obesity [21]. This could help the medical practitioner to describe the side effects to the patients being prescribed these drugs.

2.6. *Clinical trials test the drug in Humans*

Company tests drugs in actual patients on larger scale. The company has to keep track of data about patient progress. The Government wants to protect health of citizens, many rules clinical trials. In developed countries food and drug administration oversees trials. The Data mining techniques used here can be neural networks. Here data is collected by pharmaceutical company but undergoes statistical analysis to determine success of trial. Data is generally reported to food and drug administration department and inspected closely. Too many negative reactions might indicate drug is too dangerous. An adverse event might be medicine causing drowsiness.

2.7 Bioinformatics

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of
new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino
acid sequences. Development of this type of database involved not only design issues but the development
of complex interfaces whereby researchers could both access existing data as well as submit new or revised data. Ultimately, however, all of this information must be combined to form a comprehensive picture of normal cellular activities so
that researchers may study how these activities are altered in different disease states. Therefore, the field
of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. The three terms bioinformatics, computational biology and bio information infrastructure are often times used interchangeably. These three may be defined as follows:

1. Bioinformatics refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;
2. Computational biology encompasses the use of algorithmic tools to facilitate biological analysis.
3. Bio information infrastructure comprises the entire collective of information management systems, analysis tools and communication networks supporting biology.

2.7.1 Biological Database

A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of
information. For example, a record associated with a nucleotide sequence database typically contains information such as contact name, the input sequence
with a description of the type of molecule, the scientific name of the source organism from which it
was isolated, and often, literature citations associated with the sequence

2.7.2 Importance of Bioinformatics

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes: a more global perspective in experimental design the ability to capitalize on the emerging technology of database-mining – the process by which testable hypotheses are generated regarding the function or structure of a gene or.protein of interest by identifying similar sequences
in better characterized organisms

2.7.3 Scope and use of bioinformatics:

Bioinformatics is used in analyzing genomes protein bioinformatics fall into main tasks which are given below sequences, three-dimensional modelling of bio
molecules and biologic systems, etc. Different biological problems considered within the scope of

- Alignment and comparison of DNA, RNA, and protein sequences.
- Gene finding and promoter identification from DNA sequences.
- Interpretation of gene expression and microarray data.
- Gene regulatory network identification.
- Construction of phylogenetic trees for studying evolutionary relationship.
- Protein structure prediction and classification.
- Molecular design and molecular docking.

2.7.4 Sequence analysis

Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes. The sequence analysis implies subjecting a DNA or peptide sequence to sequence alignment, sequence
databases, repeated sequence searches, or other bioinformatics methods on a computer[22].

2.7.5 Genome annotation

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White[23][24]

2.7.6 Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization etc. All of these techniques are extremely noise-prone and subject to bias in the biological[25] measurement. Here the major research area involves developing statistical tools to separate signal from noise in high-throughput gene expression studies[26][27][28].

2.7.7 Modeling biological systems

Modeling biological systems is a significant task of systems biology and mathematical biology. Computational systems biology aims to develop and use efficient algorithms, data structures, visualization and communication tools for the integration of large quantities of biological data with the goal of computer modeling. It involves the use of computer simulations of biological systems, like cellular subsystems such as the networks of metabolites and
enzymes, signal transduction pathways and gene regulatory networks to both analyze and visualize the complex connections of these cellular processes[29][30].

2.7.8 Application of Data Mining in Bioinformatics

Applications of data mining to bioinformatics include 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, and protein sub-cellular location prediction. For example, microarray technologies are used to predict a patient's outcome [31][32]. On the basis of patients' genotypic [33] microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning [34] can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable [35][36][37][38].

III. METHODOLOGY AND APPROACHES

3.1 Other approaches

3.1.1 Manifold learning

- Manifold learning methods, e.g., Isomap, LLE, maintain the local patterns of distribution during transform,
- Extract features suitable for k-NN classifiers
- Can be used to reduce the dimensionality of Bio. Data.

3.2 Semi-supervised learning

- What if we have 10% labeled data, but the rest 90% are unlabelled?
- Build clusters around the labeled samples.
- Samples in the same cluster are labeled as from the same class, assuming they follow the normal distributions.

3.3 *Explainable and Accurate Data Mining Methods*

- Current methods, such as SVMs, discriminant analysis, neural networks, are 'black box' models.
- The learned knowledge is hard to understand by biologists.
- Some potential solutions
- Logic based method, e.g., decision trees and variants may be better in giving the 'IF THEN' like rules that explicitly define the epigenetic logics in cancer and stem cell development.
- DNA methylation rules can be learned by using SVM based recursive feature elimination and fuzzy logics.
- [Gene selection for cancer classification using support vector machines', Machine Learning, 2002.]

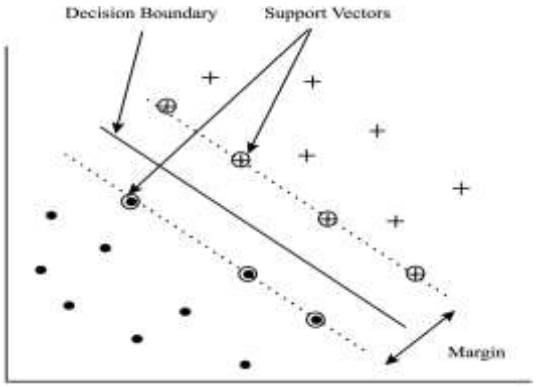

Fig:3 Data mining Analysis with Neural Network

3.4 challneges

3.4.1 Epigenetic Analysis

- Epigenetic events dominate the growth of cancer and embryonic stem cells
  - These two type of cells are of great importance
- Genes can be turned on/ off through Cytosine methylation or Histone modifications
  - The logics of DNA methylation underlie the cells' behaviors
- Wish to Know: Methylation status of CpG sites
  - CpG islands/ promoter regions in DNA sequence
  - Cancer prediction
- Traditional methods, SVMs, ANNs are
  - 'black box' models

- Knowledge are trained connection weights, or Support Vectors.
- Hard to understand for biologists

3.4.2 Adaptive Cascade Sharing Trees

- Objective: learn human understandable rules that define the epigenetic process in cancer and embryonic stem cells
- Idea:
  - Adaptively partition the numeric attributes into a set of the linguistic domains, e.g., 'high', 'very high', 'Medium', 'Low', 'Very Low'
    - Method: clustering
  - Train a committee of trees to select the most salient features and predict by voting

3.5 Method: tree learning

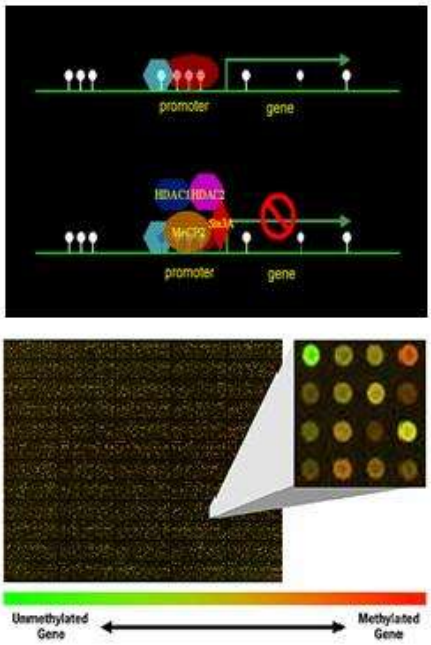

Fig:4 Data mining Approach in tree learning method

3.5 Transfer Learning

- In real life, data are hard to obtain
  - Biological experiments are expensive
  - However, biological data are related
- Can we leverage the knowledge learned in one task/domain/data set for prediction of another?
  - Humans often do this: having learned one language, find it easier to learn another
  - In Web mining, having learned to classify one web site, use the abstract knowledge to help classify another web site
- Challenge: can we leverage the knowledge learned from one data set to classify/cluster/predict another

3.6 Using Network Structure in Biology

- Adaptive Response of a Gene Network to Environmental Changes by Fitness-Induced Attractor Selection', Plos One, 2006
- The gene network is formulated as

$$\frac{d}{dx}t1 = \frac{p(A)}{1+t2^2} - w(A) * t1 + y1$$

$$\frac{d}{dx}t2 = \frac{p(A)}{1+t1^2} - w(A) * t2 + y2$$

differential equations, given some initial state the network stabilized at some attractors, corresponding to the different cell types.

- The complex dynamics of the gene networks can explain the high diversity of the species.
- Given some perturbations, how will the state of the gene networks change to adjust the levels of gene expression to environment factors?
- The dynamics of a gene network are described by differential equations, e.g., a simplified network involving only two gene nodes is formulated as:
- where m1 and m2 are the gene expression levels.
- S(A) and D(A) are the rate coefficients of synthesis and degradation. They depend on A, which represents cellular activity.
- g1 and g2 represent the noises in gene expression.

3.7 Adaptive Response of a Gene Network to Environmental Changes by Fitness-Induced

Given the initial condition, the gene expression levels stabilize at the attractors determined by the coefficients of equations. Real world gene networks can be much more complex, involving thousands of genes, leading to the complex patterns of attractors and cell activities

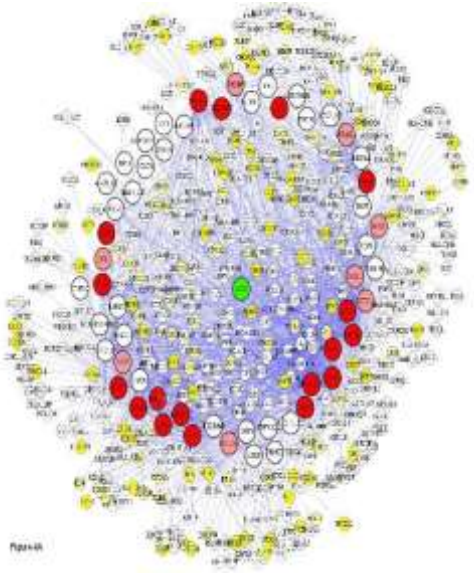

Fig:5 Data mining in *Adaptive Response of a Gene Network to Environmental Changes*

IV. CONCLUSION

This paper covers the overview of various approaches of Data mining in various medicinal and biomedical and Bioinformatics fields which is one of the interdisciplinary field of science. And it has been found during our preparation of this paper that Data mining is quit suitably for the field of bioinformatics and medical field and hence data mining has covered various faces like their size, shape, and types of data bases like biological data bases etc. Now a days data mining and biomedical filed is having of full of research aspects if somebody is tends to reach towards the research in data mining and biomedical then most of the things can be resolves and benefited for the peoples.

REFERNCES

[1]. Feelders, A., Daniels, H. and Holsheimer, M. (2000) 'Methodological and Practical Aspects of Data Mining', Information and Management, pp.271-281.
[2]. Fayyad, U.M., Piatsky Shapiro, G. and Smyth, P. (1996) From Data Mining to Knowledge Discovery in Data Base, AI Magazine, pp.37-54.
[3]. W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In Advances in Knowledge Discovery and Data Mining, pages 249–271. AAAI Press, Menlo Park, 1996.
[4]. S. Wrobel. An algorithm for multi-relational discovery of subgroups. In Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD '97, pages 78–87, London, UK, UK, 1997. Springer-Verlag.
[5]. Principles of Soft ComputinS.N.Sivanandam and S.N.Deep
[6]. K.R.Venugopal, K.G. Srinivasa and L.M. Patnaik soft computing for data Mining Application
[7]. Berthold Michael and Hand David, J. (1999) Intelligent Data Analysis: An Introduction. Springer, pp.3-10.
[8]. S. Muggleton, editor. Inductive Logic Programming. Academic Press, London, 1992.
[9]. N. Lavraˇc and S. Džeroski. Inductive Logic Programming: Techniques and Applications. Ellis Horwood New York, 1994.
[10]. S. Džeroski and N. Lavraˇc, editors. Relational Data Mining. Springer, New York, 2001
[11]. S. Kramer, N. Lavraˇc, and P. A. Flach. Propositionalization approaches to relational data mining. In N. Lavraˇc and S. Džeroski, editors, Relational Data Mining, pages 262–286. Springer, 2001.
[12]. F. Železný and N. Lavraˇc. Propositionalization-based relational subgroup discovery with RSD. Machine Learning, 62(1-2):33–63, 2006.
[13]. N. Lavraˇc, B. Kavšek, P. Flach, L. Todorovski, and S. Wrobel. Subgroup discovery with CN2-SD. Journal of Machine Learning Research, 5:153–188, 2004.
[14]. I. Trajkovski, N. Lavraˇc, and J. Tolar. SEGS: Search for enriched gene sets in microarray data. Journal of Biomedical Informatics, 41(4):588–601, 2008.
[15]. Berson A and Smith, S.J (2005) Data Warehousing, Data Mining & OLAP, Tata McGraw-Hill Edition, pp.5-6.
[16]. Prins, S and Stegwee, R. A. (2000) 'Zorgproducten en geintegreerde infromatiesystemen', (in Dutch) Handboek sturen met zorgproducten, F3100-3, december.
[17]. Morrisey, J. (1995) 'Managed care steers info systems', Modern Healthcare, Vol-25,8.
[18]. Zuckerman and Alan, M. (2006) 'Healthcare Strategic Planning', Prentice Hall of India.
[19]. Armoni,A. (2002) 'Effective Healthcare information systems', IRM Press.
[20]. Cooman De Frankey. (2005) 'Data mining in a Pharmaceutical Environment', Belgium press.
[21]. 'Novartis Business Intelligence report' Cognos press 2004 , www.cognos.com
[22]. Khalid Raza," application of data mining in bioinformatics", Indian Journal of Computer Science and Engineering, Vol 1 No 2, 114-118
[23]. Zaki , J.; Wang , T.L. and Toivonen, T.T. (2001). BIOKDD01: Workshop on Data Mining in Bioinformatics".
[24] . Li, J.; Wong, L. and Yang, Q. (2005). Data Mining in Bioinformatics, IEEE Intelligent System, IEEE Computer Society.
[25]. Liu, H.; Li, J. and Wong, L. (2005). Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data, Bioinformatics, vol. 21, no. 16, pp. 3377–3384
[26]. Yang, Qiang. Data Mining and Bioinformatics: Some Challenges, http://www.cse.ust.hk/~qyang
[27]. Berson, Alex, Smith, Stephen and Threaling, Kurt, Building Data Mining Application for CRM", Tata McGraw Hill.
[28]. Zhang, Yanqing; C., Jagath, Rajapakse, Machine Learning in Bioinformatics, Wiley, ISBN: 978-0-470-11662-3
[29]. Kuonen, Diego. Challenges in Bioinformatics for Statistical Data Miner, Bulletin of the Swiss Statistical Society, 46; 10-17.
[30]. Richard, R.J. A. and Sriraam, N. (2005). A Feasibility Study of Challenges and Opportunities in Computational Biology: A Malaysian Perspective, American Journal of Applied Sciences 2 (9): 1296-1300.
[31]. Tang, Haixu and Kim, Sun. Bioinformatics: mining the massive data from high throughput genomics experiments, analysis of biological data: a soft computing approach, edited by Sanghamitra Bandyopadhyay, Indian Statistical Institute, India
[32]. Nayeem, Akbar; Sitkoff, Doree, and Krystek, Jr., Stanley. (2006) A comparative study of available software for highaccuracy homology modeling: From sequence alignments to structural models, Protein Sci. April; 15(4): 808–824
[33]. N., Cristianini and M., Hahn. (2006) Introduction to Computational Genomics, Cambridge University Press. ISBN 0-5216-7191-4.
[34]. SJ, Wodak and Janin, J. (1978). Computer Analysis of Protein-Protein Interactions. Journal of Molecular Biology 124 (2):323–42.
[35]. Mewes, H.W.; Frishman, D.; X.Mayer, K. F.; Munsterkotter, M., Noubibou , O.; Pagel, P. and Rattei, T. (2006) Nucleic Acids Research, 34, D169.
[36]. Lee, Kyoungrim. (2008). Computational Study for Protein-Protein Docking Using Global Optimization and Empirical Potentials, Int. J. Mol. Sci. 9, 65-77.
[37]. Hirschman, Lynette; C. Park, Jong; T., Junichi, Wong, L. and H. Wu., Cathy (2002). Accomplishments and

challenges in literature data mining for biology, BIOINFORMATICS REVIEW, Vol. 18 no. 12, 1553–1561

[38]. 1Mohammed Waseem Ashfaque;2Abdul Samad Shaikh; 3Sumegh Tharewal; 4Sayyada Sara Banu; 5Mohammed Ali Sohail, Challenges of Interactive Multimedia Data Mining in Social and Behavioral Studies for latest Computing &Communication of an Ideal Applications, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. VII (Nov – Dec. 2014), PP 21-31.