www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 8 Aug 2016, Page No. 17499-17505

A Review on Clinical Data Mining

Taranath N.L.[1] Dr. Shantakumar b Patil.[2] Dr. Premajyothi Patil.[3] Dr. C.K. Subbaraya[4]
Research Scholar,
Dept. of CS& E, NCET,
VTU
taranath.taras@gmail.com
Prof. & Head, Dept. of CS& E,.
NCET, Bengaluru.
Shantakumar.p@gmail.com
Dept. of CS& E, NCET,
Bengaluru.
Prof. & Prinicipal, Dept. of CS& E,.
AIT, Chikkamagaluru.
subrayack@gmail.com

Abstract :

Clinical data mining is a practice based research strategy by which practitioners and researchers retrieve, analyze and interpret available qualitative and quantitative information from available medical records. It is an active interdisciplinary area of research that is considered the consequent of applying artificial intelligence and data mining concepts to the field of medicine and health care. The aim of this work is to provide a review on the foundation principles of mining clinical dataset, and present the findings and results of past researches on utilizing data mining techniques to mine health care data and patient records. The scope of this article is to present a brief report on preceding investigations made in the sphere of mining clinical data, the techniques applied and the conclusions recounted. The most recent research findings that can further unveil the potential of data mining in the realm of health care and medicine are clearly presented in this review.

Keywords : Artificial Intelligence, Data Mining, Clinical data mining, Health care.

1. Introduction

Data mining concepts [1] are focused on discovering knowledge, predicting trends and eradicating superfluous data. Data is available in enormous magnitude, but the knowledge that can be inferred from the data is still negligible [2]. Discovering knowledge [5] in medical systems and health care scenarios is a herculean yet critical task. Knowledge discovery [2] [3] describes the process of automatically searching large volumes of data for patterns that can be considered additional knowledge about the data [4]. The knowledge obtained through the process may become additional data that can be used for further manipulation and discovery [3].Application of data mining concepts to the medical arena has undeniably made remarkable strides in the sphere of medical research and clinical practice saving time, money and life [5-9]. Clinical data mining is the application of data mining techniques using clinical data [7]. Clinical Data-Mining (CDM) involves the conceptualization, extraction, analysis, and interpretation of available clinical data for practical knowledge-building, clinical decision-making and practitioner reflection [9]. The main objective of clinical data mining is to haul new and previously unknown clinical solutions and patterns to aid the clinicians in diagnosis, prognosis and therapy[8][9][10]. Moreover application of software solutions to store patient records in an electronic form is expected to make mining knowledge from clinical data less stressful [11].

There is a growing need in the health care scenario to store and organize sizeable clinical data, analyze the data, assist the health care professionals in decision making, and develop data mining methodologies to mine hidden patterns and discover new knowledge from clinical data[4][11]. The basic steps involved in clinical data mining include data sampling, data analysis, data modernization, data modeling and data ranking[6][7][10]. The focus of this research is to explore and present an overview of the fundamental models and frameworks of mining clinical data, investigate existing results of mining patient records of varied nature, and brief about the challenges encountered in mining patient records.

2. Literature review on CDM

There have been a great number of surveys and studies in the area of data mining, and each of the phases in data mining viz, Clustering, Feature selection, Outlier Detection and Classification play a major role in unearthing significant clinical patterns from patient records and inferring previously unknown knowledge [12][14][15]. The following sections present a brief survey on previous and recent reviews on mining clinical data.

2.1 Clinical Data Mining

Hanauer [12] reported the challenges and solutions in mining electronic data for research and patient care. The Michigan Health system statistics were utilized for their

research. However the author was concerned and focused on the hurdles involved in text mining alone. The challenges that the author inferred included affirmation of accurate diagnosis and natural language processing of electronic health records. The author had provided a solution called EMERSE (Electronic Medical Record Search Engine) that provided keyword searches for basic users and advanced features for power users. The interface was user-friendly, secure and compliant with privacy regulations and practical for implementation. However the system needed more training and the searching procedures continued to raise complexity. Roddick et.al, [11] presented the experiences of the authors in applying exploratory data mining techniques to medical health and clinical data. This enabled the authors to elicit a number of general issues and provided pointers to possible areas of future research in data mining and knowledge discovery from a broad perspective. Iavindrasana et.al [7], used the nine data mining steps proposed by Fayyad in 1996 [8] as the main themes of the review. MEDLINE [16] was used as the primary source and 84 papers were retained by the authors for analysis. Their results identified three main objectives of data mining that were stated as follows: understanding of the clinical data, providing assistance to healthcare professionals, and formulating a data analysis methodology to explore clinical data. Classification was stated to be the most frequently used data mining function with a predominance of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines). A myriad of quantitative performance measures were proposed with a predominance of accuracy, sensitivity, specificity, and ROC curves. Further work was reported by Lalayants et.al [17] who described a practice-based, mixed-method research methodology stating Clinical Data-Mining (CDM) to be a strategy for engaging international practitioners for describing, evaluating and ruminating upon endogenous forms of practice with the ultimate goal of improving practice and contributing to knowledge[9]. Such knowledge contributions were considered to be localized, but through conceptual reflection with empirical replication they could be generalized. A more elaborate account on mining clinical data was done by Epstein and Irwin as stated in the following lines. Epstein and Irwin [9] provided a clear definition of clinical data mining (CDM) as a practice-based, retrospective research strategy whereby practitioner-researchers alone or with the assistance of a research consultant, systematically extracted, codified, investigated and interpreted available qualitative and quantitative data from their own and other agency records in order to reflect on the practice, program and policy implications of their findings. Methodologically, they identified three types of clinical studies namely quantitative data, narrative data and qualitative data. Their work included a number of CDM research descriptions spanning an assortment of clinical populations covering pediatric diabetes, adolescent mental health, domestic violence, liver transplantation, geriatrics, and palliative care across the lifespan [18][19][20]. According to their survey, CDM was considered advantageous in terms of cost savings in using existing data within practices compared to creating data for prospective analysis. However they also brought about the existing flaws in CDM viz, missing data and discrepancies amongst workers on work description and record maintenance that could significantly affect and reduce the reliability and validity of practitioner-generated information.

The above reviews have focused on the broad domain of mining clinical data, the challenges encountered and the results inferred from clinical data analysis. The ensuing sections deal with specific techniques to mine and disinter patterns, associations and clusters of medical records. Moreover methods to manage and analyze voluminous patient records have also been reviewed in this work.

3. Data Mining Models in CDM

Clinical data mining analysis crafts effective and worthwhile knowledge that is indispensable for precise and accurate decision making [21]. Various types of mining models have been used in the past to represent interesting facts and latent patterns and trends in clinical datasets with copious applications in medical practice [22] [23]. In this subsection some of the data mining models applied to healthcare are briefly reviewed.

3.1 Feature Relevance Models :

Clinical data are generally voluminous in nature and need special attention by virtue of data storage and analysis. Feature relevance analysis[24][25] is a phase in data mining that enables researchers to filter out certain predictors of ailments from further exploration under the pretext of being less contributory to the detection of an ailment[26]. For instance, a patient's health record may contain the concerned Patient ID, Address, and Occupation along with the evidenced clinical findings and laboratory investigation results among other details. The former factors are highly inessential in diagnosing the patient's state of health and time spent on analysis of such details is a huge squander. Such attributes need to be filtered out from further analysis and this would certainly save time and lessen computational complexity.

3.2 Clustering Models

Clustering is derived from mathematics, statistics, and numerical analysis [27] [28]. In this technique the dataset is partitioned into two or more factions (clusters) of similar records [29]. The clustering algorithms aim at grouping records keeping in mind the ultimate objective of maximizing a similarity metric between the members of the cluster [30]. In most cases, closeness is the similarity metric and the aim is to maximize the cumulative closeness between data records in a cluster [29] [30]. The researchers then explore the properties of the members of the generated clusters.

3.3 Association Models

Association rule(X) Y is defined over a set of transactions T where X and Y are sets of items. In a Clinical setting, the set T can be patients' clinical records and items may be symptoms, measurements, observations, or diagnosis corresponding to the patients' clinical records. Given S as a set of items, support(S) is defined as the number of transactions in T that contain all members of the set S. The confidence of a rule (X) Y is defined as support(X(Y)/support(X)), and the support of this rule is support(X(Y)). The discovered association rules show

hidden patterns in the mined dataset. For example, the rule: ({People who are alcoholic})/ {People needing dialysis} with a high confidence signifies that the number of people requiring dialysis is high among people who are alcoholic.
The following sections are devoted to a review on past work on clinical data investigated through data mining techniques and models.

3.4 Clinical Data Mining Frameworks

The general approach to mine clinical data comprises of the following phases namely Data collection, Data Pre-processing, Feature Selection, Classification and Evaluation [32]. Inclusion of Outlier detection prior to Classification could reduce computational complexity and remove sparse and unrelated patient data. We also attempt to summarize the Clustering techniques to group similar medical records into classes. Moreover the dependencies among symptoms and diseases can also be identified through Association Rule Mining.
Abe et.al, [28] introduced the concept of categorized and integrated data mining. The authors reviewed the rapid progress in medical science, medical diagnosis and treatment and perceived the need for an integrated and cooperative research among medical researchers, biology, engineering, cultural science, and sociology. Hence they proposed a framework called Cyber Integrated Medical Infrastructure (CIMI), a framework of integrated management of medical data on computer networks consisting of a database, a knowledge base, and an inference and learning component, connected to each other in the network. The framework had the capacity to deal with diverse types of data which required integrated analysis of diverse data. In their study, for medical science, they analyzed the features and relationships among various types of data and revealed the possibility of categorized and integrated data mining. Anand et.al, [19] presented a framework to incorporate parallelism in mining data. This led to the enhancement of algorithms being developed within this framework to be parallel and was hence expected to be efficient for large data sets, a definite need for medical data mining. The parallelism within the framework permitted distribution and heterogeneity. The framework could be easily updated and new discovery methods could be readily incorporated within the framework. The framework provided a spontaneous view of handling missing data during the discovery process using the concept of Ignorance borrowed from Evidence Theory. The framework incorporated the possibility of representing data and knowledge, and methods for data manipulation and knowledge discovery. They suggested an extension of the conventional definition of mass functions in Evidence Theory for use in Data Mining, as a means to represent evidence of the existence of rules in the database. The discovery process within EDM consisted of a series of operations on the mass functions. Each operation was carried out by an EDM operator. Also included was a classification for the EDM operators based on the discovery functions performed by them and a discussion of induction, domain and combination operator classes was carried out. The application of EDM to two separate Data Mining tasks was also addressed, highlighting the advantages of using a general framework for Data Mining and, in particular, using one that was based on Evidence Theory. Lin and Haug, [27] proposed an approach to data preparation that utilized information from the data, metadata and sources of medical knowledge. Heuristic rules and policies were defined for the three types of supporting information. Compared with an entirely manual process for data preparation, their proposed approach could potentially reduce manual work by achieving a degree of automation in the rule creation and execution. A pilot experiment demonstrated that data sets created through the approach lead to better model learning results than a fully manual process. This study was conducted using data extracted from the enterprise data warehouse (EDW) of Intermountain Health Care (IHC) in Salt Lake City. The data was captured during routine clinical care documented in the HELP [4] hospital information system. IHC had established a working process that duplicated data from the HELP system to a data mart in the EDW called the "HELP" data repository. The authors developed a system to detect patients who were admitted to the hospital with pneumonia. The data set included data of patients who were discharged from the hospital with pneumonia as primary diagnosis as well as a group of control patients. Both groups were sampled from patients admitted to LDS Hospital from the year 2000 to 2004. The manual approach was to acquire variables relevant to pneumonia according to domain knowledge and the medical literature. Keyword searches were used on the code description field to find a list of candidate data codes. The candidate codes were inspected and the most suitable codes were chosen. The earliest observed value for each code was selected as the summary value for the chosen period. Each time no value was found for an instance of a variable, the variable was discretized and a state called 'missing' was added to it. By using the aforementioned process, a data set in flattened-table format was created from the original data. In their experimental approach, two types of heuristic rules were used to select variables. One was to pre-screen data elements based on their statistical characteristics and their gross categorization in the data dictionary. This allowed selection of data subsets that were relevant to the clinical model being developed. The second was to select data elements that were able to differentiate the specific clinical problem according to comparative statistics calculated from the test and control groups across all candidate variables. The candidate variable list was then manually inspected to remove obviously irrelevant variables. The numbers of patients in the case and control groups were 1521 and 1376 respectively. The two 95% confidence intervals of the difference of ROC were found to be above zero, indicating that the difference was statistically significant (a=0.05). The results revealed the fact that the two tested model learning algorithms performed better with the data set prepared by the framework.
Kazemzadeh et.al, [22] focused on encoding, sharing, and using the results of data mining analyses for clinical decision making at the point of care. With the aforesaid objective in mind, a knowledge management framework was proposed that addressed the issues of data and knowledge interoperability by adopting healthcare and data mining modeling standards, HL7 and PMML respectively. A prototype tool was developed as part of their research that provided an environment for clinical guideline authoring

and execution capable of applying and interpreting data mining results. Moreover, three real world case studies were presented. The authors also described a novel framework for dissemination and application of the data and mined knowledge among the heterogeneous healthcare information systems. For data interoperability, HL7 Clinical Document Architecture (CDA) schema was used to define the required structure for encoding patients' health related data. The healthcare researchers extracted knowledge by mining existing healthcare data in an off-line operation and stored in proprietary databases. They used the PMML specification to encode the produced mined knowledge to achieve knowledge interoperability between sources of knowledge and their users. This was reported to be the first methodology to make this type of knowledge portable and available at the application sites. Further on, decision modules could access patient data from CDA documents and supply them into the data mining models from the PMML documents. The results of this operation were also provided as CDA documents to allow interoperability of the results. Moreover, the authors utilized the mined knowledge in the proposed extension to the GLIF3 clinical guideline modeling language that provided recommendations and warnings to the healthcare personnel based on the results of knowledge application.

Rapid and extensive research has attracted science and engineering professionals to work in a cohesive manner to the advancement in the domain of clinical data mining. The following section depicts a brief outlook on the rising spheres of CDM.

4. Applications of Clinical Data Mining

Several reviews and surveys have been reported in the past that have portrayed the impact of data mining techniques in refining health care applications[5][6][8][9]. A concise view of the recent work in the area of clinical data mining and their contribution to the advancement of clinical practice, data management and research is presented in this section.

4.1 Data Mining in Clinical Data Management

Data pre-processing techniques have been widely used in management of medical data and patient records [14] [15]. The large volume of data available needs to be formatted and collected in a manner that will permit secure and simple retrieval when needed, faster and efficient mining of credential information and economic utilization of storage space and computation time. Electronic health records [12] [14] were the first attempt to securely manage the patient records and are currently used in practice in several medical institutions and health care centers around the world. Distributed network of medical records was another innovation that spurred a renaissance in the medical field that allowed clinicians to share patient information for the purpose of obtaining an expert opinion or sharing the storage space available in another network and even for providing backup facility. Mining of information from such distributed records is currently an intense area of research.

Knowledge-Based Systems/Clinical Decision Support Systems Several studies have been reported on the results of mining medical data by application of data mining techniques that include feature selection, outlier detection and classification/ prediction. Each of the algorithms is evaluated and the technique that produces the best classification accuracy is chosen. The rules generated by the classification algorithm and the medical data records on which the data mining techniques were executed constitute the Knowledge Base which is the core component of any data mining framework. Following this any medical record relative to the particular ailment under study can be input to the classifier and the precision in classification can be verified from the clinical decision of the system. Hence such classifier systems offer support to the medical practitioners in predicting the course of a disease based on the existing symptoms, proposing drugs, identifying the need for hospitalization and predicting possible time for recuperation. Other data mining applications related to clinical practice include associating the various side-effects of treatment, collating common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub-populations that respond differently from the mainstream population to certain drugs, and determining proactive steps that can reduce the risk of affliction.

Data mining techniques thus provide better assessment of patient needs, better information about clinical fidelity, superior idea about patient outcomes and association between interventions and outcomes[4][26]. This has stimulated application of computational techniques in mining of medical data streams, customer relationship management and fraud detection related to non-compliance with security/ethical issues of clinical data.

4.2 DNA Sequence Analysis for Genetic Marker Detection

Data mining techniques have proven to produce improvement in the analysis, classification of the affection status of more individuals and by locating more single nucleotide polymorphisms related to the disease. Molecular genetic markers represent one of the most influential tools for the analysis of genomes and enable the association of inborn traits with underlying genomic diversity. Molecular marker technology has developed rapidly over the last decade and two forms of sequence based markers, Simple Sequence Repeats (SSRs), also known as microsatellites, and Single Nucleotide Polymorphisms (SNPs) now preponderate applications in modern genetic analysis [88]. The diminishing price of DNA sequencing has led to the availability of large sequence data sets derived from whole genome sequencing and large scale Expressed Sequence Tag (EST) discovery have enabled the mining of SSRs and SNPs. These can later be applied to diversity analysis, genetic trait mapping, association studies, and marker assisted selection. These markers are economical, require minimal labour to produce and can frequently be associated with annotated genes.

In recent years there has been an upsurge in the rate of acquisition of biomedical data. Advances in molecular genetics technologies, such as DNA microarrays allow laymen for the first time to obtain a comprehensive view of the cell. Machine learning and statistical techniques applied to gene expression data have been used to focus on the questions of distinguishing tumour morphology, predicting post-treatment outcome, and finding molecular markers for

diseases. On record today, the microarray-based classification of different morphologies, lineages and cell histologies can be performed successfully in many instances. The performance in predicting treatment outcome or drug response has been quite limited although some results are quite promising. Most results of microarray analysis still require further experimental validation and follow up study. In a few cases the results of microarray analysis have found their way into more serious consideration in clinical use. Recent advances in data mining applications include Gene Selection that is a process of attribute selection, which finds the genes most strongly related to a particular class, Clustering aims at finding new biological classes or refining existing ones and Classification aims at classifying diseases or predicting outcomes based on gene expression patterns, and includes identifying the best treatment for a given genetic signature. The influence exerted by data mining techniques can span wider avenues only when the current obstacles in mining medical data are handled in an appropriate manner. The following section narrates the existing challenges in mining clinical data and patient records.

5. Challenges In Clinical Data Mining

Clinical data mining is certainly limited by the ease of access to medical findings, since required facts for data mining often exist in different settings, forms and systems, viz, administration, clinics, laboratories and other. This calls for a strategy to gather and integrate data before data mining can be done. While several authors and researchers have suggested the need for a data warehouse prior to mining clinical data, the expenses involved challenge their utility. However, Intermountain Health Care have successfully implemented a warehouse from five different sources-a clinical data repository, acute care case-mix system, laboratory information system, ambulatory case-mix system, and health plans database and imparted better evidence-based clinical solutions. Research by Oakley [33] suggested a distributed network topology instead of a data warehouse for more efficient data mining. Another imposing hurdle in medical data collection include missing, distorted, conflicting, and non-homogenous data, such as bits of information recorded in different formats in diverse data sources. Precisely, the lack of a standardized clinical vocabulary is a serious hindrance to data mining. Cios and Moore [34] have posed a dispute that data problems in healthcare are the result of the dimensionality, intricacy and assorted nature of medical data and their low mathematical characterization and non conformance to a certain protocol. Moreover ethical, legal and social issues encountered in CDM also have to be appropriately handled. The issue of obtaining patterns of diverse nature on exhaustive mining of data needs to be deliberated upon. Extensive research may reveal many interesting patterns and relationships not necessarily valuable. The successful application of data mining requires expertise in data mining methodology and tools not ignoring realistic knowledge of medical practice. Data mining applications in healthcare can have tremendous potential and efficacy. However, the success of healthcare data mining hinges on the availability of clean healthcare data [4-6] [17] [23]. In this respect, it is critical that the healthcare industry consider diverse ways and means of capturing, storing, processing and mining data. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. Further, as healthcare data are not limited to patient records, it is necessary to explore the use of text and image mining approaches to expand the scope and nature of clinical data mining.

6. Conclusion

Data mining is one of the extensively researched areas in computer science and information technology owing to the wide influence exhibited by this computational technique on diverse fields that include finance, clinical research, multimedia, education and the like. CDM is a highly motivated area of research due to the extensive influence exerted by this multi-domain research area that brings together interests of medical practitioners, computer science researchers and health care professionals. Mining of clinical facts is highly essential due to the availability of exhaustive and enormous volume of medical records. This paper presented a review of clinical data mining concepts and the data mining techniques applied in clinical practice. Designs of Data mining framework for clinical data mining systems have been reviewed to provide researchers an initiative to formulate new techniques for clinical record analysis and exploration, besides reforming the flaws in the existing systems. Also stated are the principles behind the existing applications of clinical data mining, the challenges existing in CDM and future directions for research. This research study is expected to be a significant contribution to researchers and practitioners in the data mining and clinical industry.

7. References

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques" (3 Ed.). Elsevier. ISBN 978-0-12-374856-0

[2] Cabena, Peter, Pablo Hadjnian, Rolf Stadler, Jaap Verhees and Alessandro Zanasi (1997). "Discovering Data Mining: From Concept to Implementation" Prentice Hall, ISBN 0-13-743980-6.

[3] Xingquan Zhu, Ian Davidson (2007). "Knowledge Discovery and Data Mining: Challenges and Realities." Hershey, New York. p. 18. ISBN 978-1-59904-252-7.

[4] Debahuti Mishra , Asit Kumar Das, Mausumi and Sashikala Mishra, "Predictive Data Mining: Promising Future and Applications", Int. J. of Computer and Communication Technology, Vol. 2, No. 1, 2010

[5] Dave Smith, SAS, Marlow, UK, "Data Mining in the Clinical Research Environment", PhUSE 2007.

[6] Prasanna Desikan, Hsu, Srivastava, "Data mining for health care management", 2011 SIAM International Conference on Data mining.

[7] Iavindrasana J et.al, Clinical data mining: a review. Med Inform. 2009:121-33. Review.

[8] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. Retrieved 2008-12-17.

[9] Epstein, Irwin. (2010). Clinical data-mining: Integrating practice and research. London. Oxford University Press

[10] Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005). Introduction to Data Mining. ISBN 0-321-32136-7

[11] John F.Roddick, Peter Fule, Warwick J.Graco, "Exploratory Medical Knowledge Discovery: Experiences and Issues", 2004.

[12] David Hanauer, MD, MS Mining clinical electronic data for research and patient care: Challenges and solutions, Clinical Assistant Professor University of Michigan, USA, 2007 September

[13] R. Agrawal et al., Fast discovery of association rules, in Advances in knowledge discovery and data mining pp. 307–328, MIT Press, 1996.

[14] Bennett CC and TW Doub. (2010) "Data mining and electronic health records: Selecting optimal clinical treatments in practice". Proceedings of the 6th International Conference on Data Mining. pp. 313-318.

[15] M.F. Ochs et al. (eds.), "Clinical Research Systems and Integration with Medical Systems", Biomedical Informatics for Cancer Research,DOI 10.1007/978-1-4419-5714-6_2, © Springer Science Business Media, LLC 2010

[16] Medline Resources http://www.nlm.nih.gov/bsd/pmresources.html

[17] Lalayants et.al, "Clinical data-mining: Learning from practice in international settings", International Social Work March 27, 2012, doi: 0020872811435370

[18] Jerome Beker, Anthony J Grasso Dsw, Irwin Epstein, Boysville Of Michigan, Information Systems in Child, Youth, and Family Agencies, Published October 11th 1993 by CRC Press.

[19] Irwin Epstein, Susan Blumenfield, Clinical Data-Mining in Practice-Based Research, May 7th 2002 by Routledge

[20] Irwin Epstein, Ken Peake, Daniel Medeiros, Clinical and Research Uses of an Adolescent Mental Health Intake Questionnaire, August 14th 2005 by Routledge

[21] Gregory Piatetsky-Shapiro, Pablo Tamayo, "Microarray Data Mining: Facing the Challenges" SIGKDD Explorations. Volume 5, Issue 2.

[22] Weiss and Indurkhya. Predictive Data Mining. Morgan Kaufmann

[23] Riccardo Bellazzi, Blaz Zupanb, Predictive data mining in clinical medicine: Current issues and guidelines"", international journal of medical informatics 77 (2008) 81–97.

[24] G. Bontempi. "Structural feature selection for wrapper methods". In Proceedings of ESANN 2005, European Symposium on Artificial Neural Networks, 2005.

[25] Jiang et.al, Feature Mining Paradigms for Scientific Data, Copyright © by SIAM

[26] Archana Venkataraman, Marek Kubicki, Carl-Fredrik Westin, Polina Golland, "Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies", 978-1-4244-7028-0/10/$26.00 ©2010 IEEE

[27] M. Sacha. (2008) "Clustering of a periodical medical Knowledge Constrained K-means Clustering with Background data." in Proceedings of the Eighteenth International Conference on Machine Learning, 2001, a periodical-medical-data. pp. 577 - 584.

[28] G. Y. Hang, D. Zhang, J. Ren, and C. Hu, "A Machine Learning Repository: Hierarchical Clustering Algorithm Based on K-Means with Constraints," in Fourth International Conference on Innovative Computing, Information and Control, Kaohsiung, Taiwan, 2009, pp. 1479-1482

[29] Lin W. and C. Le "Model-based cluster analysis of microarray gene expression data". Genome Biology, 3(2): research0009.1-0009.8, (2002).

[30] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6, November 2010

[31] V.Elango, R.Subramanian,V.Vasudevan, "A Five Step Procedure for Outlier Analysis in Data Mining", European Journal of Scientific Research, ISSN 1450-216X Vol.75 No.3 (2012), pp. 327-339.

[32] Berner, E. Clinical decision support systems: theory and practice. 2007. Springer Verlag.