ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

A Survey on the principles of mining Clinical Datasets by utilizing Data mining technique

[1]S. Sharath, [2]M. N. Rao,[3]H. G. Chetan

[1]P. G. Student, Dept. of CS&E, A.I.T., Visvesvaraya Technological University,Chikmagalur, India

[2]P. G. Student, Dept. of CS&E, M.C.E.,Visvesvaraya Technological University,Hassan, India

[3]P. G. Student, Dept. of CS&E, A.I.T., Visvesvaraya Technological University,Chikmagalur, India

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. This paper presents survey on the foundation principles of mining clinical datasets by utilizing data mining techniques to mine health care data and patient records. It also focuses on the preceding survey made in clinical data mining, techniques applied and the conclusion drawn. The most recent research findings that can further unveil the potential of data mining in the realm of healthcare and medicine are clearly presented in this survey.

KEYWORDS: Artificial intelligence, Data Mining, Clinical data mining, Clinical Datasets, Health care, Patient Records.

I. 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, analyse 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 modelling 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.

II. RELATED WORK

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][15].

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.

2.2*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.

2.2.1*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.

2.2.2*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.

2.2.3*Outlier Detection Models*

Outlier detection models signify novelty, anomaly, noise, variation or could be categorized under mining exceptions [31]. Definition derived from Barnette & Lewis (1994) stated that an outlying observation, or outlier, is one that appeared to deviate markedly from other members of the sample in which it occurs. Indicated in study, outlier normally being considered as noise, and recently under data mining approaches outliers were considered as significant details to drill out important information. One of the steps towards obtaining a coherent analysis is detection of outlying observations [34].Detection of extreme observations could eradicate incorrect data while at the same time presence of Outliers could lead to novel insights in clinical knowledge discovery. Hence Outliers pose a challenge in the domain of CDM and need to be handled appropriately.

2.2.4*Classification Models*

A classification algorithm assigns a patient’s data record with specific attributes and attribute- values to a predefined class. The classification techniques in healthcare are generally applied for diagnostic purpose. A classification model is built using a set of relevant attribute-values (records) derived from clinical facts and findings that lead us to generate different categories representing different nature of records. On comparison of a new patient’s record with those of patients in different classes, one can determine to which class the new patient belongs, for instance a benign class (Non-Cancerous) or a malignant one (Cancerous).

2.2.5*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 patient’s 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.

2.3*Data Mining Techniques in CDM*

This survey is aimed at providing a review on past researches in the domain of mining clinical datasets comprising of patient records and clinical findings. Data mining techniques commonly applied in the medical domain aim at classification of disease nature or prediction of the course of an ailment. Clustering and Association rule mining have been utilized in cases where similar patient records and related symptoms needed investigation [31]. Early work on CDM was reported by Prather et.al, [18] who stated that clinical databases tend to accumulate large quantities of information about patients and their medical conditions. Venkataraman et.al [26], proposed an alternative to Univariate statistics to detect population differences in functional connectivity. The authors proposed a feature selection method based on a procedure that could search across subsets of the data to isolate a set of robust and predictive functional connections. A metric called Gini Importance was introduced that could summarize multivariate patterns of interaction, that could not be captured by Univariate techniques.Three feature selection techniques, the stepwise feature selection (SFS), sequential floating forward search (SFFS), and principal component analysis (PCA), and two commonly used classifiers, Fisher's linear discriminant analysis (LDA) and support vector machine (SVM), were investigated. Samples were drawn from multidimensional feature spaces of multivariate Gaussian distributions with equal or unequal covariance matrices and unequal means, estimated from a clinical data set. Classifier performance was quantified by the area under the receiver operating characteristic (ROC) curve Az. The LDA and SVM with radial kernel performed similarly for most of the conditions evaluated in their study.PCA was comparable to or better than SFS and SFFS for LDA at small samples sizes, but inferior to SVM with polynomial kernel. For the class distributions simulated from clinical data, PCA did not show advantages over the other two feature selection methods. Under this condition, the SVM with radial kernel performed better than the LDA when few training samples were available, while LDA performed better when a large number of training samples were available. Chih Lee et.al, [25] investigated the Linear Discriminant Analysis, Sequential Probability Ratio Test(SPRT) and a modified SPRT (MSPRT) empirically using clinical datasets from Parkinson’s disease, colon cancer, and breast cancer. The authors assumed the same normality assumption as LDA and proposed variants of the two SPRT algorithms based on the order in which the components of an instance were sampled. Leave-one-out cross-validation was used to assess and compare the performance of the methods. Their results indicated that two variants, SPRT-ordered and MSPRT-ordered, were superior to LDA in terms of prediction accuracy. Moreover, on average SPRT-ordered and MSPRT-ordered examined fewer components than LDA before arriving at a decision. The reported results suggested that SPRT-ordered and MSPRT-ordered were the preferred algorithms over LDA.Jacob and Ramani, [30] addressed the task of grading the performance of feature selection and classification algorithms on clinical datasets of varied nature. In 2011, the authors investigated the performance of sixteen data mining classification algorithms viz. Random Tree, Quinlan’s decision tree algorithm

(C4.5), K-Nearest Neighbour algorithm etc., on the Wisconsin Breast tissue dataset (derived from the UCI Machine Learning Repository) that comprised of 11 attributes and 106 patient records. The analysis indicated the level of training accuracy and other performance measures of the algorithms in detecting the presence of breast cancer and the associated breast tissue conditions that raised the risk of developing cancer in future.

Jacob et.al, [23] [12] further explored the performance of classification algorithms on the Breast Cancer dataset through Data mining algorithms. Their research aimed at recognizing the significance of feature selection in classifying Breast Cancer data under two classes namely Benign (non-cancerous) and Malignant (cancerous). They examined the performance of six feature relevance algorithms on the Wisconsin Breast Cancer (WBC) dataset comprising of 699 patient records and the Wisconsin Diagnostic Breast Cancer (WDBC) dataset comprising of 569 patient details obtained from the UCI Machine Learning Repository. A comparison of twenty classification algorithms based on their Misclassification Rate was portrayed. It is also stated here that the C4.5 algorithm offered more efficient classification since the decision tree size generated comprised of fewer nodes than the Random Tree algorithm.

Jacob et.al, [7] [13] analysed the performance of twenty classification algorithms on the (Oncovirus) cancerous Hepatitis C virus dataset from the UCI Machine Learning Repository. The authors performed binary classification on the dataset, comprising of 155 instances and 19 predictor features. The performance of the classification algorithms revealed that Random Tree Classification and Quinlan's C4.5 algorithm classified the patient records as infected and healthy with 100% accuracy. The Quinlan's C4.5 algorithm and the Random Tree classification algorithm revealed 99.97% classification accuracy on the DNA data. In a another major attempt to explore the classification algorithms, Jacob executed classification algorithms on diverse clinical data comprising of patient records related to the following ailments viz, Mammography masses, Heart Disease, Dermatology infection, Orthopaedic ailment and Thyroid diseases. The authors made a careful selection of clinical data from varied domains in order to identify the performance of the data mining algorithms on different types of clinical datasets.

Jacobanalysed the Cardiotocography dataset from the UCI Irvine Machine Learning Repository comprising of 2126 Fetal Heart Rate (FHR) and morphology pattern (MP) records and 21 predictor attributes providing life saving information on the state of the fetus in the womb. They classified the fetal records into three target classes and their analysis state 100 percent classifier accuracy for Random Tree and Quinlan's C4.5 algorithm in the case of the FHR dataset.Some classical examples for inward procedures of Outlier detection can be found in [22]. In the medical scenario, this may lead to neglecting or overlooking patients suffering from a rare disease or exhibiting a rare combination of symptoms. To account for this conditional aspect of outlier detection in medicine Chauhan et.al, [30] made a detailed study of Hierarchical, Partitioning-based, Density-based and Grid-based clustering algorithms. Their conclusions suggested the use of K-Means and Hierarchical Agglomerative Clustering for mining clinical databases. Wilson et al. [36] discussed potential uses of data mining techniques in pharmaco-vigilance to detect adverse drug reactions. Ramesh Kumar [14] had proposed an algorithm named nVApriori to mine interesting rules from HIV infected Patient's Treatment records. This is a n-cross validation based Apriori (nVApriori) algorithm to mine domain irrelevant rules. Acquired Immune Deficiency Syndrome (AIDS) is a critical disease in the medical domain. The author proposed a new dataset for AIDS/HIV infected patients' case history.Exarchos devised a new automated methodology based on Association rules for detection of Ischemic beats in long duration Electrocardiographic recordings (ECG). The proposed approach comprised of three stages viz, Pre-processing, Discretization and Classification. Noise removal and extraction of required features was done during the Pre-processing phase. According to the proposed methodology, electrocardiogram (ECG) features extracted from the ST segment and the T-wave, as well as the patient's age, was used as inputs. The obtained sensitivity (Se) and specificity (Sp) was 87% and 93%, respectively.Vararuk et.al, [23] made use of data mining techniques to extract and investigate patterns in HIV/AIDS patient data. These patterns aimed to provide better management of the disease and targeting of resources. A total of 250,000 anonymised records from HIV/AIDS patients in Thailand were imported into a database. IBM's Intelligent Miner was used for clustering and association rule discovery. The significance of the study was stated as follows: Identification of symptoms that were precursors of other symptoms could allow the targeting of the former so that the later symptoms could be avoided. The research showed provisioning a pragmatic and targeted approach to the management of resources available for HIV/AIDS treatment. The authors work suggested implementation of a quality monitoring system to target available resources.Ordonez et al. [32] proposed a new algorithm to mine association rules in medical data with additional constraints on the extracted rules and applied the method to predict heart disease. A decision tree-based classification approach was applied to mass spectral data to help diagnosis of ovarian cancer suspects.

III. CLINICAL DATA MINING SYSTEMS

Clinical data mining [7] refers to the collection of algorithms, techniques and methods to discover previously unknown, new patterns from clinical data that could aid clinicians, heath-care practitioners, medical researchers, and scientists in disease diagnosis and prognosis, genetic marker detection and drug therapy. The basis for any data mining framework involves a preliminary learning phase during which the problem is modeled followed by the test phase that validates the constructed model. The learning process can be accomplished either in a Supervised or Unsupervised manner [1]. Supervised Learning [10] requires the training data to be accompanied by class labels and the test data is classified based on the training set, whereas in unsupervised learning, the class label is unknown and the aim is to establish the existence of clusters or classes in the data. Models required to mine data are classified into Predictive and Descriptive models. Clustering and Association Rule Models are descriptive while Classification and Regression models are stated to be predictive.

3.1*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] [19]. 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. The parallelism within the framework permitted distribution and heterogeneity. 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. 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.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. The second was to select data elements that were able to differentiate the specific clinical problem. 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.

Mc. Gregor et.al, [14] [20] presented a framework for process mining in critical care. The CRoss Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996, with the goal of being industry, tool and application-neutral. Constant references to the methodology by analysts have established it as the de facto standard for data mining and Knowledge Discovery in Databases (KDD). The proposed framework utilized the CRISP-DM model, extended to incorporate temporal and multidimensional aspects (CRISP-TDMn), in conjunction with the Patient Journey Modeling Architecture (PaJMa), to provide a structured approach to knowledge discovery of new condition onset pathophysiology in physiological data streams. A portray of the instantiation of the framework for late onset neonatal sepsis was given, using CRISP-TDMn for the process mining model and PaJMa for the knowledge representation. Their research presented a generalized framework to support process mining in critical care that enabled knowledge discovery of new condition onset pathophysiology using temporal data mining of physiological data streams and constructed process flow mappings that could be used to update patient journeys as instantiated within clinical practice guidelines.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. For data interoperability, HL7 Clinical Document Architecture (CDA) schema was used to define the required structure for encoding patients' health related data. 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. 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.

IV. 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][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 here.

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] 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.

4.2*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 [4][26].

4.3*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. 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. 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.

V. 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 efficiency. However, the success of healthcare data mining hinges on the availability of clean healthcare data [4-6] [17] [23]. 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.

VI. CONCLUSION AND FUTURE SCOPE

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 survey 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. 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. This research study is expected to be a significant contribution to researchers and practitioners in the data mining and clinical industry.

REFERENCES.

1. Ian H. Witten, Eibe Frank and Mark A. Hall, “Data Mining: Practical Machine Learning Tools and Techniques”, Elsevier. ISBN 978-0-12-374856-0, 3rdEdition, pp. 313-317, 2000.
2. Cabena, Peter, Pablo Hadjnian, Rolf Stadler, Jaap Verhees and Alessandro Zanasi, “Discovering Data Mining: From Concept to Implementation”, Prentice Hall, ISBN 0-13-743980-6, pp. 97-99, 1997.
3. Xingquan Zhu and Ian Davidson, “Knowledge Discovery and Data Mining: Challenges and Realities”, International Conference on Data mining, Hershey, New York, ISBN 978-1-59904-252-7, pp.14-18, 2007.
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, Issue No. 1, 2010.
5. Dave Smith and Marlow, “Data Mining in the Clinical Research Environment”, PhUSE, pp. 89-94, 2007.
6. Prasanna Desikan, Hsu and Srivastava, “Data mining for health care management”, 2011 SIAM International Conference on Data miningpp.24-28, 2011.
7. Iavindrasana J et.al, “Clinical data mining: a review”, Med Information, pp. 121-133,2009.
8. Fayyad, Usama, Gregory Piatetsky-Shapiro and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases",http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf, 1996.
9. Epstein and Irwin, “Clinical data-mining: Integrating practice and research”, London, Oxford University, Press 2010.
10. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, “Introduction to Data Mining”, London, Oxford University, Press 2005.
11. John F.Roddick, Peter Fule and Warwick J.Graco, “Exploratory Medical Knowledge Discovery: Experiences and Issues”,PhUSE, pp. 69-74, 2004.
12. David Hanauer, “Mining clinical electronic data for research and patient care: Challenges and solutions”, Clinical Assistant Professor, University of Michigan, USA, September2007.
13. R. Agrawal et al., “Fast discovery of association rules, in Advances in knowledge discovery and data mining”, MIT Press, pp. 307–328, 1996.
14. Bennett CC and TW Doub, “Data mining and electronic health records: Selecting optimal clinical treatments in practice”, Proceedings of the 6th International Conference on Data Mining, pp. 313-318, 2010.

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, doi: 0020872811435370, March 27, 2012.
18. Jerome Beker, Anthony J Grasso Dsw and Irwin Epstein, Boysville Of Michigan, “Information Systems in Child, Youth, and Family Agencies”, Published by CRC Press, October 11th 1993.
19. Irwin Epstein and Susan Blumenfield, “Clinical Data-Mining in Practice-Based Research”, Routledge,May 7th 2002.
20. Irwin Epstein, Ken Peake and Daniel Medeiros, “Clinical and Research Uses of an Adolescent Mental Health Intake Questionnaire”, Routledge August 14th 2005.
21. Gregory Piatetsky-Shapiro, Pablo Tamayo, “Microarray Data Mining: Facing the Challenges”, SIGKDD Explorations, Volume 5, Issue No 2, 2011.
22. Weiss and Indurkhya, “Predictive Data Mining”, Morgan Kaufmann, 2006.
23. Riccardo Bellazzi and Blaz Zupanb, “Predictive data mining in clinical medicine: Current issues and guidelines”, International journal of medical informatics, Vol 7, pp. 81–97, 2008.
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, 2005.
26. Archana Venkataraman, Marek Kubicki, Carl-Fredrik Westin and Polina Golland, “Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies”, IEEE 9768-1-4244-7028-0/10/$26.00 ©2010, 2010.
27. M. Sacha, “Clustering of a periodical medical Knowledge Constrained K-means Clustering with Background data”, In Proceedings of the Eighteenth International Conference on Machine Learning, a periodical-medical-data. pp. 577 – 584, 2001.
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, pp. 1479-1482, 2009.
29. Lin W. and C. Le “Model-based cluster analysis of microarray gene expression data”, Genome Biology, Vol 2, Issue 3, 2002.
30. Ritu Chauhan, Harleen Kaur and M.Afshar Alam, “Data Clustering Method for Discovering Clusters in Spatial Cancer Databases”, International Journal of Computer Applications (0975 – 8887), Volume 10, Issue No.6, November 2010.
31. V.Elango, R.Subramanian and V.Vasudevan, “A Five Step Procedure for Outlier Analysis in Data Mining”,European Journal of Scientific Research, ISSN 1450-216, Vol.75, Issue No.3, pp. 327-339, 2012.
32. Berner E., “Clinical decision support systems: theory and practice”, Springer Verlag,2007.
33. T. Gunnar and A. Aamodt, “Towards an Introspective Architecture for Meta-level Reasoning in Clinical Decision Support Systems”, European Journal of Scientific Research, Vol.65, Issue No.2, pp. 227-239, 2010.
34. A. E. Smith, C.D. Nugent and S. I. McClean, “Evaluation of inherent performance of intelligent medical decision support systems: utilizing neural networks as an example”,Artificial Intelligence in Medicine,Vol.6, pp. 1-27, 2003.

BIOGRAPHY

S. Sharathis currently pursuing M.Tech in the Department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikmagalur. He received B.E. degree in 2012 from Adichunchanagiri Institute of Technology, Chikmagalur. His research interests are Data Mining and Warehousing, Artificial Intelligence, Machine learning, Computer Networks etc. He has undertaken Hybrid Medical Decision Support Systems as final year M.Tech. Project.

M. N. Rao is currently pursuing M.Tech in the Department of Computer Science and Engineering, Malnad College of Engineering, Hassan. She received B.E. degree from Atria Institute of Technology, Bangalore. Her research interests are Data Mining and Warehousing, Artificial Intelligence, Computer Networks etc. She has undertaken Electronic Patient Consent Management Systems as M.Tech. Project.

H. G. Chetanis currently pursuing M.Tech in the Department of Computer Science and Engineering, Adichunchanagiri Institute of Technology, Chikmagalur. He received B.E. degree in 2012 from SDMIT, Ujire. His research interests are Data Mining and Image Processing. He has undertaken Feedback session based user search goals prediction as final year M.Tech. Project.