Data Mining Methods and Techniques for Clinical Decision Support Systems

B. Senthil Kumar
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Anima.P
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India.

Abstract – The impact of data mining have huge growth in Health care data, the patient electronic health record analysis and decision making became very essential for all types of clinical applications. It tremendously helps to the medical experts to make a decision from the patient's electronic health records. In such scenario, the Clinical Decision Support System in health care domain has a huge impact, which utilizes the machine learning techniques and patients electronic health records to find appropriate decisions. The clinical decision support system involves the patient specific health state measurements in the analysis that provides knowledge and personalized decisions, intelligently processed and presented at appropriate times. This paper provides a review of different techniques and methods for clinical decision support system with its merits and demerits. Finally, the conclusion is given based on the review process.

Index Terms – Data Mining, Decision Support System (DSS), Machine Learning, Electronic Health Record (HER), Fuzzy, Neural Network, Genetic Algorithms.

1. INTRODUCTION

The medical domain, a vast amount of knowledge is required even to solve seemingly simple problems. A physician is required to remember and apply knowledge of a vast array of documented disease presentations, diagnostic parameters, the combination of drug therapies and guidelines. However, the physician's cognitive abilities are restricted due to factors like multitasking, limited reasoning, and memory capacity. Consequently, it is impossible for an unaided physician to make the right decision every time. Ironically, the increasing rate of information generated by medical advances has aggravated the physician's task. The ideal decision-making process contains the process of knowledge discovery. Many researchers [1] [2] consider data mining programs as a way to make decision-making tools intelligent. The potential of computer based tools to address the medical decision-making problems are realized half a century ago and several algorithms have been developed to construct Clinical Decision Support Systems for a variety of medical applications.

Clinical Decision Support System (C_DSS) is interactive application, which performs automatic decision making for clinical activities. This is designed to assist physicians and other health professionals with decision-making tasks by determining the diagnosis of patient data [3]. Fig 1.0 shows the Clinical decision C_DSS helps to review and filter the physician's preliminary diagnostic choices to improve the treatments. The Post diagnosis in C_DSS systems is used to mine data to derive associations between patients and their past medical history. Using the patient medical history and clinical research, the application will predict future events. Clinical decision support systems are broadly classified into two main types namely Knowledge based C_DSS and Non-knowledge based C_DSS, which relies on machine learning approaches [4].

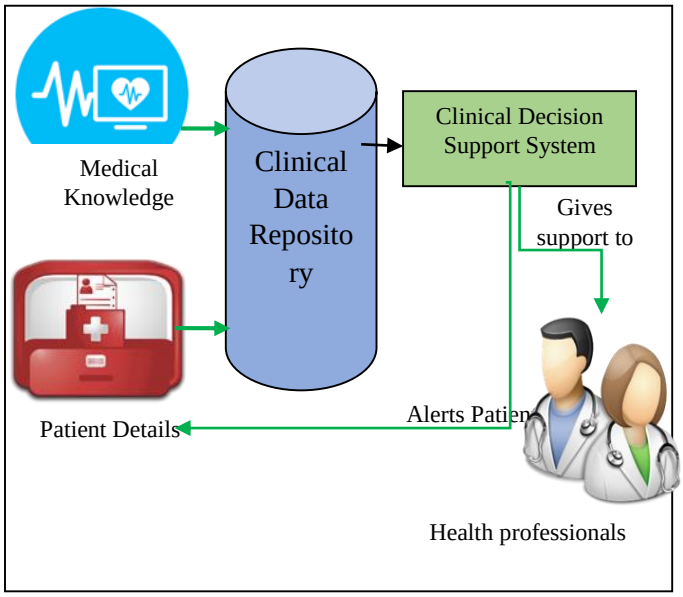

Fig 1.0 Clinical decision support architecture

1.1. Knowledge Based C_DSS:

The knowledge based clinical decision support system consists of several rules. The rules are in the form of if-then statements, which is generated from the knowledge base. The knowledge based generally consists of three main parts, one is Knowledge based, and second one is Inference rules and finally a mechanism to communicate. It is used to show the result to the users as well as to provide input to the system for effective

decision making. There are the different kinds of C_DSS available in the medical health centers. The knowledge based systems are commonly have some particular task in the rule generation. The knowledge within the expert system is generally represented as the set of rules. Sometimes the knowledge based is used with variance management to execute patient care process and provide high-quality health care services dynamically. Knowledge based C_DSSs are then further divided into three main categories. They are Fuzzy Logic Rule Based C_DSS, Bayesian Network, Rule-based systems, and Evidence based Systems.

1.1.1 Fuzzy Logic Rule Based: Fuzzy logic is a knowledge based approach that facilitates solutions to resolve vagueness in decision support system .Fuzzy logic rule based approach can be a very helpful for relating abstraction and imprecision in precise mathematical language, explicitly representing clinical ambiguity. The Fuzzy Logic Rule based classifier is very effective in the high degree of positive predictive value and diagnostic accuracy.

1.1.2 Rule- Based Systems & Evidence Based Systems: The rule and evidence based systems are tend to capture the knowledge of domain experts into expressions that can be evaluated as rules. When a large number of rules have been compiled into a rule based system then the working knowledge will be evaluated against this technique. This performs the rule aggregation until a result is attained. It is helpful for storing a large amount of data and information. However, it is difficult for an expert to transfer their knowledge into distinct rules. For closing the gap between the physicians and C_DSSs, evidence based appeared to be a perfect technique. It proves to be a very powerful tool for improving clinical care and also patient outcomes. It has the potential way o improve in terms of quality and safety as well as reducing the cost.

1.1.3 The Bayesian network: The Bayesian network is the knowledge based graphical representation that shows a set of variables and their probabilistic relationships between diseases and symptoms. They are based on conditional probabilities, the probability of an event given the occurrence of another event, such as the interpretation of diagnostic tests. Baye's rule helps us compute the probability of an event with the help of some more readily available information and it consistently processes options as new evidence is presented. In the context of C_DSS, the Bayesian network can be used to compute the probabilities of the presence of the possible diseases given their symptoms. This type of techniques has numerous merits, that include the knowledge and conclusions of experts in based on the probabilities that are applicable to many models. The demerits of Bayesian Network include the difficulty to get the probability knowledge for possible diagnosis and not being practical for large complex systems given multiple symptoms. The Bayesian calculations on multiple simultaneous symptoms could be overwhelming for users.

1.2 Non-Knowledge Based C_DSS:

C_DSS without a knowledge base is called as non-knowledge based C_DSS. These Systems instead used a form of artificial intelligence called as machine learning. Non-knowledge based C_DSSs are then further divided into four main categories. They are Artificial Neural Networks, Genetic Algorithms, statistical methods and Hybrid systems.

1.2.1 Artificial Neural Network: Artificial Neural Networks (ANN) is non-knowledge based adaptive C_DSS that uses a form of artificial intelligence, also known as machine learning, that allows the systems to learn from past experiences and recognizes patterns in clinical information. It consists of nodes called neuron and weighted connections that transmit signals between the neurons in a forward or looped fashion. An ANN consists of 3 main layers: Input (data receiver or findings), Output (communicates results or possible diseases) and Hidden (processes data). The system becomes more efficient with known results for large amounts of data. The advantages of ANN include the elimination of needing to program the systems and providing input from experts. The ANN C_DSS can process incomplete data by making educated guesses about missing data and improves with every use due to its adaptive system learning. Some of the disadvantages are that the training process may be time-consuming leading users to not make use of the systems effectively. The ANN systems derive their own formulas for weighting and combining data based on the statistical recognition patterns over time which may be difficult to interpret. Neural Networks have been widely applied to nonlinear statistical modeling problem and for modeling large and complex databases of medical information. The goal of training is to optimize the performance of the network in estimating output for particular input space. Back propagation training algorithm, a popular approach used with medical databases adjusts the weight of an ANN to minimize a cost function. The ANN maintains correct classification rates and allows a large reduction in complexity of the systems.

1.2.2 Genetic Algorithms: A Genetic Algorithm is a non-knowledge based method based on Darwin's evolutionary theories that dealt with the survival of the fittest. These algorithms rearrange to form different combinations that are better than the previous solutions. Similar to neural networks, the genetic algorithms derive their information from patient data. An advantage of genetic algorithms is these systems go through an iterative process to produce an optimal solution. The fitness function determines the good solutions and the solutions that can be eliminated. A disadvantage is the lack of transparency in the reasoning involved for the decision support systems making it undesirable for physicians. The main challenge in using genetic algorithms is in defining the fitness criteria. In order to use a genetic algorithm, there must be many components such as multiple drugs, symptoms, treatment therapy and so on available in order to solve a problem.

1.2.3 Statistical Method: It is one of most simple and useful method used for data collection. It can be in the form of a survey, experiment result or questionnaire. Development of clinical decision support systems using the statistical method as an integral part is very common. Data can be collected as a questionnaire mentioning the status of patients how he looks like, their way of talking, what he feels and much more.

1.2.4 Hybrid Systems: A combination of two or more methodologies within a design of single system results into a hybrid system. Hybrid systems extract the best from all methodologies and provide an optimal solution for clinical decision support systems. Meta reasoning method such as hybrid systems consists of different reasoning methodologies. It can consist of a rule based, case based and model based reasoning. That finally results in an overall improvement of the system performance.

2. LITERATURE SURVEY

In the paper [5] authors developed the framework of web services along with the Bayesian theorem to construct an SOA based Medical Decision Support System to help medical experts, this gives an appropriate decision for the issues related to medical diagnosis process. The heterogeneous Medical Decision Support System enhanced the accuracy, quality, and efficiency of medical diagnosis and offered more service platforms than the conventional one.

In the paper [6] authors explored a Health Maintenance Facility (HMF) to provide the equipment and supplies necessary to deliver medical care to the Space Station. The essential part of the HMF was a computerized Medical Decision Support System (MDSS) that enhanced the ability of the medical officer (paramedic or physician) to maintain the crew's health and provided the emergency medical care.

Authors in [7] presented a weighted fuzzy rule-based C_DSS for the risk prediction of heart patients consisting of two phases: (1) automated approach for the generation of weighted fuzzy rules and (2) developing a fuzzy rule-based DSS. In the first phase, the authors used the mining technique, attribute selection and attribute weight method to obtain the weighted fuzzy rules. Then, the fuzzy system was constructed in accordance with the weighted fuzzy rules and chosen attributes. The experiments were carried out on using the datasets obtained from the UCI repository and the performance was compared with the neural network-based system utilizing accuracy, sensitivity and specificity.

Authors in [8], have developed an expert system approach based on Principal Component Analysis and Adaptive Neuro-Fuzzy Inference System for diagnosis of diabetes disease. The aim of the study was to improve the diagnostic accuracy of diabetes disease combining PCA and ANFIS. The proposed system has two stages. In the first stage, the dimension of diabetes disease data set that has features is reduced to 4 features using Principal Component Analysis. In the second stage, diagnosis of diabetes disease is conducted via Adaptive NeuroFuzzy Inference System classifier. The diabetes disease dataset used in this study was from the UCI Machine Learning Database. The obtained classification accuracy of the system was 89.47% and it was very promising with regard to the other classification applications in literature for this problem.

Authors in [9] introduced a new strategy for design and development of an efficient dynamic DSS for supporting rare cancers decision making that operated on a Graphics Processing Unit (GPU) and was capable of adjusting its design in real time based on user-defined clinical questions in contrast to standard CPU implementations that were limited by processing and memory constraints using a Probabilistic Neural Network classifier. The proposed GPU-based DSS was evaluated on 140 rare brain cancer cases and its ability to predict tumors' malignancy achieved 78.6% overall accuracy.

In paper [10], authors have presented insights from clinical data repositories of patient data sets. Clinical repositories containing large amounts of biological, clinical, and administrative data are increasingly becoming available as health care systems integrate patient information for research and utilization objectives. To investigate the potential value of searching these databases for novel insights, they have applied a new data mining approach, HealthMiner, to a large cohort of 667,000 inpatient and outpatient digital records from an academic medical system. HealthMiner approaches knowledge discovery using three unsupervised methods: CliniMiner, Predictive Analysis, and Pattern Discovery. The initial results from this study suggest that these approaches have the potential to expand research capabilities through identification of potentially novel clinical disease associations.

Authors in [11] evaluated the pre-/post-implementation effect of a C_DSS on the performance of prospective audit with intervention and feedback and demonstrated an increase in interventions and recommendation acceptance countered by a substantial number of non-actionable alerts. The authors suggested that C_DSSs for antimicrobial stewardship required considerable human resources and financial investments.

In paper [12], authors examined the use of a new guideline based Computerized Clinical Decision Support system for asthma in a pediatric pulmonology clinic of a large academic medical center that included patterns of computer use in relation to patient care, and themes surrounded the relationship between asthma care and computer use. The authors found that Pediatric pulmonologists demonstrated the low use of a computerized DSS for asthma care because of a combination of general and subspecialist-specific factors.

Authors in [13] focused on medical diagnosis by learning pattern through the collected data of diabetes, hepatitis and heart diseases and developed intelligent medical DSS to help

the physicians. The authors proposed the use of decision trees C4.5 algorithm, ID3 algorithm and CART algorithm to classify these diseases and compared the effectiveness, correction rate among them.

Authors in [14] designed knowledge-based systems in medicine for diagnostic tasks by applying Fuzzy set theory and fuzzy logic and verified the approach with trials with the systems; A fuzzy expert system for syndromes differentiation in oriental traditional medicine, an expert system for lung diseases using fuzzy logic, case based reasoning for medical diagnosis using fuzzy set theory, a diagnostic system combining disease diagnosis of western medicine with syndrome differentiation of oriental traditional medicine, a fuzzy system for classification of western and eastern medications and a fuzzy system for diagnosis and treatment of integrated western and eastern medicine. Results from experiments showed that fuzzy logic was known to resolve the conflicts aroused from ambiguity, uncertainty, and imprecision of information.

Authors in [15] evaluated the accuracy of a computerized C_DSS designed to support assessment and management of pediatric asthma in a subspecialty clinic. The results showed that computerized C_DSS performed relatively accurately compared to clinicians for assessment of asthma control but was inaccurate for treatment. Authors also presented a hybrid approach based on feature selection, fuzzy weighted pre-processing and Artificial Immune Recognition System (AIRS) to medical DSSs for heart and hepatitis diseases datasets, taken from UCI machine learning database. AIRS showed an effective performance on machine learning benchmark problems and medical classification problems like breast cancer, diabetes, and liver disorders classification, and obtained classification accuracies of 92.59% and 81.82% using 50- 50% training-test split for heart disease and hepatitis disease datasets, respectively. Hence, the proposed method can be used in medical DSSs.

In paper [16], authors developed a DSS using artificial intelligence techniques for the classification and diagnosis of epilepsies and epilepsy syndromes in children by importing the clinical and laboratory data that achieved diagnoses overall success rate of 93.4%. The authors suggested that the system was helpful for trainees, decision making and differential diagnosis.

Authors in, [17] a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser).He later presented C4.5 (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared. Author later published the book Classification and Regression Trees (CART), which described the generation of binary decision trees. ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples.

Bayes' theorem is used for clinical decision support system, which is named as Thomas Bayes after modified by the author in [18], who did early work in probability and decision theory. Bayesian classification is based on Bayes' theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. In this paper, author proved Bayesian classifiers have exhibited high accuracy and speed when applied to large databases.

Authors in [19] have proposed k-nearest neighbor algorithm (k-NN) for classifying objects based on closest training examples in the feature space. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms in which an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors.

In paper [20], authors have presented a back propagation algorithm which performs learning on a multilayer feed-forward neural network. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer. Back propagation learns by iteratively processing a data set of training tuples, associations of the network's prediction for each tuple with the actual known target value, this learning mechanism identifies the accurate class label at the time of classification. The target value may be the known class label of the training tuple for the classification problems.

Authors proposed Support Vector Machines, a promising method for the classification of both linear and nonlinear data in [21]. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyper-plane using support vectors which are the essential training tuples and margins that are defined by the support vectors. From the literature, several algorithms and techniques are analyzed.

Authors in [22] [23] surveyed the various techniques and methods used for medical data classification and finally provided the summary about the algorithms. So along with the decision support system, the classification problems are also identified.

Many Clinical decision supports systems are reviewed in the section II, However, the work in the earlier study was not concentrated on the personalization in C_DSS. That does not considered the explicit characteristics of patients for recommendations and those techniques and methods are

limited to some type of applications and not much accurate in the results.

3. CONCLUSION

This paper presented a detailed survey on the different techniques and methods used in the design of a Clinical Decision Support Systems. The conclusion from this paper is that the implementation of proper knowledge based DSS system can improve the detection accuracy. Along with the C-DSS approaches, the paper concludes that the Genetic Algorithm can be used for feature selection in clinical data sets and it can find the optimal solution more quickly and improve the efficiency of the C_DSS. The survey exposed that the Artificial Neural Networks (ANN) and fuzzy classification rules using data mining techniques can incorporate data from many clinical and laboratory variables to provide better diagnostic accuracy in various clinical datasets. By keeping all the above concluding remarks of this paper, design and development of efficient feature selection and classification techniques for Clinical Decision Support Systems is found as a challenging issue in the research. So deploying optimal methods for complete machine learning techniques with the real time application oriented issue is suggested in the future work.

REFERENCES

[1] Bara, Adela, and Ion Lungu. "Improving decision support systems with data mining techniques." *Advances in Data Mining Knowledge Discovery and Applications*. InTech, 2012.

[2] Burn-Thornton, Kath E., and Simon I. Thorpe. "Improving clinical decision support using data mining techniques." *AeroSense'99*. International Society for Optics and Photonics, 1999.

[3] Musen, Mark A., Blackford Middleton, and Robert A. Greenes. "Clinical decision-support systems." *Biomedical informatics*. Springer London, 2014. 643-674.

[4] Kawamoto, Kensaku, et al. "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success." *Bmj* 330.7494 (2005): 765.

[5] Chang, C. C. and H. M. Lu, (2009), "A SOA-Based Medical Diagnosis Decision Support System using the Bayesian theorem and web service technology", Journal of the Chinese Institute of Engineers, 32(7), 923-930.

[6] Ostler, D. V., R. M. Gardner and J. S. Logan, (1988), "A Medical Decision Support System for the Space Station Health Maintenance Facility", Proceedings of the Annual Symposium on Computer Application in Medical Care, 43-47

[7] Anooj, P. K. "Implementing decision tree fuzzy rules in clinical decision support system after comparing with fuzzy based and neural network based systems." *IT Convergence and Security (ICITCS), 2013 International Conference on*. IEEE, 2013.

[8] Polat, Kemal, and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease." *Digital Signal Processing* 17.4 (2007): 702-710.

[9] Sidiropoulos, K., Glotsos, D., Kostopoulos, S., Ravazoula, P., Kalatzis, I., Cavouras, D., & Stonham, J. (2012). Real time decision support system for diagnosis of rare cancers, trained in parallel, on a graphics processing unit. *Computers in biology and medicine*, *42*(4), 376-386.

[10] Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., ... & Rigoutsos, I. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in biology and medicine*, *36*(12), 1351-1377.

[11] Hermsen, E. D., VanSchooneveld, T. C., Sayles, H., & Rupp, M. E. (2012). Implementation of a clinical decision support system for antimicrobial stewardship. *Infection Control & Hospital Epidemiology*, *33*(4), 412-415.

[12] Lomotan, E. A., Hoeksema, L. J., Edmonds, D. E., Ramírez-Garnica, G., Shiffman, R. N., & Horwitz, L. I. (2012). Evaluating the use of a computerized clinical decision support system for asthma by pediatric pulmonologists. International journal of medical informatics, 81(3), 157-165.

[13] Kumar, D. Senthil, G. Sathyadevi, and S. Sivanesh. "Decision support system for medical diagnosis using data mining." *International Journal of Computer Science Issues* 8.3 (2011): 147-153.

[14] Omidiora, E. O., M. O. Olaniyan, and O. Derikoma. "A Decision Support System Model A Decision Support System Model for Diagnosing Tropical Diseases Using Fuzzy Logic Diagnosing Tropical Diseases Using Fuzzy Logic." (2011).

[15] Lomotan, E. A., Hoeksema, L. J., Edmonds, D. E., Ramírez-Garnica, G., Shiffman, R. N., & Horwitz, L. I. (2012). Evaluating the use of a computerized clinical decision support system for asthma by pediatric pulmonologists. *International journal of medical informatics*, *81*(3), 157-165.

[16] Vassilakis, Kostas M., Lagia Vorgia, and Sifis Micheloyannis. "Decision support system for classification of epilepsies in childhood." *Journal of child neurology* 17.5 (2002): 357-362.

[17] Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1.1 (1986): 81-106.

[18] Saaty, Thomas L., and Luis G. Vargas. "Diagnosis with dependent symptoms: Bayes theorem and the analytic hierarchy process." *Operations Research*46.4 (1998): 491-502.

[19] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." *IEEE transactions on information theory* 13.1 (1967): 21-27.

[20] Stuart Russell and Peter Norvig, "Artificial Intelligence: A Modern Approach," Prentice Hall, 2010.

[21] Bernhard E. Boser, Isabelle Guyon and Vladimir Vapnik, "A Training Algorithm for Optimal Margin Classifiers," Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, pages: 144-152, 1992.

[22] Senthil Kumar, B & Gunavathi R., Dr. (2016). A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis. IJARCCE. 5. 463-467. 10.17148/IJARCCE.2016.512105.

[23] Senthil Kumar, B & Sreejith.R. (2016)."A Survey on Identification of Diabetes Risk Using Machine Learning Approaches". IJIRCCE. 4. 33 -335 . 10.15680/IJIRCCE.2016. 0408001.