SHIV SHAKTI
International Journal in Multidisciplinary and Academic Research (SSIJMAR)
Vol. 3, No. 1, February- March -2014 (ISSN 2278 – 5973)

A Review on Data Mining Techniques Used in Healthcare Industry

Dimple*

Abstract

Data mining is a growing area of research that complements with many disciplines such as Artificial Intelligence (AI), databases, statistics, visualization, and high-performance and parallel computing[44]. The goal of data mining is to turn data that are facts, numbers, or text which can be processed by a computer into information andknowledge[8]. Nowadays, the health care relies on data very much. Therefore, this paper aims to understand about data mining and its importance in medical systems.The goal of studying data mining techniques for the diagnosis and prognosis of various diseases is to identify the well-performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees,Artificial neural networks and their Multilayer Perceptron model, Naïve Bayes. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the algorithms mentioned above together results more effective.

Keywords:Data Mining, Decision Tree (DT), Artificial Neural Network (ANN), Naïve Bayes, Healthcare Database, Diagnosis.

*pre-phD student,CSE Dept., U.I.E.T., MaharshiDayanand University Rohtak

Abstract—Data mining is a growing area of research that complements with many disciplines such as Artificial Intelligence (AI), databases, statistics, visualization, and high-performance and parallel computing[44]. The goal of data mining is to turn data that are facts, numbers, or text which can be processed by a computer into information andknowledge[8]. Nowadays, the health care relies on data very much. Therefore, this paper aims to understand about data mining and its importance in medical systems.The goal of studying data mining techniques for the diagnosis and prognosis of various diseases is to identify the well-performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees,Artificial neural networks and their Multilayer Perceptron model, Naïve Bayes. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the algorithms mentioned above together results more effective.

Keywords:Data Mining, Decision Tree (DT), Artificial Neural Network (ANN), Naïve Bayes, Healthcare Database, Diagnosis.

1. Introduction

Data mining is the process of analyzing and summarizing data from different perspectives and converting it into useful information. In a large database, data mining is used to find out patterns to extract hidden pieces of information [2][46]. Data mining is defined as "a process of nontrivial extraction of implicit, previouslyunknown and potentially useful information from the data stored in a database" [1][47].The data mining processes include formulating ahypothesis, collecting data, performing preprocessing,estimating the model, and interpreting the model and draw theconclusions.

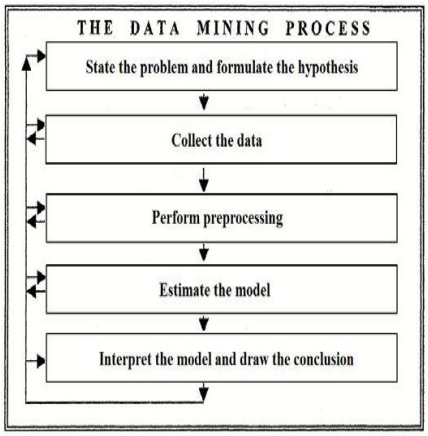


Fig.1 The Data Mining Process [3]

Healthcare databases have a huge amount of data but however, there is a lack of effective analysis tools to discover the hidden knowledge [11].Appropriate computer-based information and/or decision support systems can help physicians in their work. Efficient and accurate implementation of an automated system needs a comparative study of various techniques available[10]. In this paper an overview of the current research being carried out using the DM techniques for the diagnosis and prognosis of various diseases[53]. The rest of this paper is organized as follows: First different data mining techniques that can be used to classify the data is explained then identify

the most used algorithms for disease diagnosis and prognosis, and finally conclusions is shown.

2. Data mining technique used for classification of data

2.1 Neural Networks

An artificial neural network (ANN), often just called a "neural network" (NN), is a algorithmic based model or mathematical and computational model based on biological neural network[5][48]. An (ANN) artificial neural network, also called a neural network, is a mathematical model based on biological neural networks [9]. A neural network consists of an interconnected group of artificial neurons. Neural networks are used to model complex relationships between inputs and outputs or to find patterns in data.

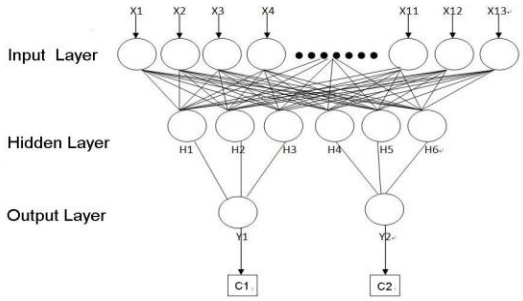


Fig2: Framework of Neural network containing three layers [9]

It maps a set of input data onto a set of appropriate output data [4] .It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of

$Yj = f(\Sigma\ wji\ xi)$

2.2 Decision Tree

The decision tree is a powerful classification algorithm that is popular in the information systems [9]. The decision tree is performed with separate recursive observation in branches to construct a tree for prediction. The splitting algorithms – i.e. Information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-squared test (used in CHAID) – are used to identify a variable and the corresponding threshold, and then split the input observation into two or more subgroups [9][50]. The steps are repeated until a complete tree is built as

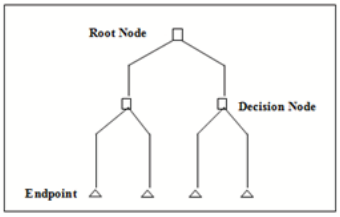


Fig.3: A Decision Tree [3]

2.3 Naïve Bayes

In probability theory, Bayes' theorem (often called Bayes' law after Thomas Bayes) relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations [5][51]. For example, a patient may be observed to have certain symptoms. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation.

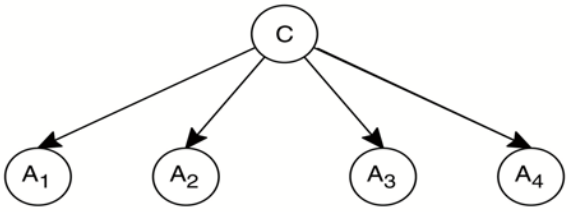


Fig.4: A Representation of a Bayesian Classifier Structure [21].

The structural model is represented as a directed graph where the nodes represent attributes and arcs represent attribute dependency.

A Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be a tomato if it is red in color, round in shape, and about 3" in diameter. This classifier takes all these features to contribute independently to the probability that this fruit is a tomato, whether or not they're in fact related to each other or to the existence of the other features.

The Bayes theorem is as follows:

Let X={x1, x2,.....,xn} be a set of n attributes. In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C[10]. To determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as

P (H|X) = P (X| H) P (H) / P (X)

As Naïve Bayes classifiers depends on the precise nature of the probability model , so it can be trained very efficiently in a supervised learning setting [3]. Here independent variables are considered for the purpose of prediction or occurrence of the event. It has been shown that Naïve Bayes classifiers often work much better in many complex real world situations [6].

An advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

2.4 Support Vector Machine

The Support Vector Machine (SVM) is a classification algorithm in statistical learning theory [13]. It can provide accurate models because it can capture nonlinearity in the data. The classification tasks are performed by maximizing the margin separating both classes and minimizing the classification errors [13]. The training of SVM involves the optimization of a convex cost function where the learning process is not complicated by local minima [14]. The testing used the support vectors to classify a test dataset and the performance is based on error rate determination [14]. For a training set of l samples, the learning procedures are as the followings [15]:

$$\min_{\alpha} : \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j .$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, i = 1, \cdots, l .$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 .$$

The yi is the label of the ith sample xi [15]. The ai is the Langrangian multiplier of xi. The C is the upper bound of ai and K(xi,xj) is the kernel. The samples with a > 0 are

called support vectors [12]. The decision function is as follow, where ns is the number of support vectors [15] :

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n_s} y_i \alpha_i^* K(x_i, x) + b^*\right).$$

3.Literature Review

There are different kinds of studies for DM techniques in medical databases. We identify the following categories:

1. Studies that summarize reviews and challenges in mining medical data in general [16], [24], [25], [31], [32]
2. Studies of DM techniques used for diagnosing and/or prognosing of specific diseases, which can be further classified into three other categories: those which use DM techniques for disease diagnosis [3],[7],[9],[14],[22],[37],[27] for disease prognosis [4],[10],[26],[29],[42],[43],[45] or both diagnosis and prognosis.[13],[36]
3. Studies to investigate factors which have higher prevalence of the risk of a disease [5],[12],[28]
4. Studies that present new technologies and algorithms [18-21], [40], [41] and studies that present new techniques improving old ones, such as [8],[11],[30],[39]
5. Studies that present new frameworks, tool and applications in medicine and healthcare system [2],[15-17],[23],[33-35],[38]

4.Application Of Data Mining In Medical Systems

The reliance of health care on data is increasing[16].Medical researchers, physicians, and health care providers face the problem to use stored data efficiently when more medical information systems with large database are used [16]. The medical information system databases contain many data such as patient records, physician diagnosis, and monitoring information where the data has been useful in many medical decision support systems to save lives [15].

A medical decision support systems are systems that help in the decision making process in the medical domains such Clinical Decision Support Systems (CDSS), medical imaging, and Bioinformatics [15]. The contributions of these systems are to reduce medical errors and costs, earlier disease detection, and to achieve preventive medicine [15]. The advantages of using computerized CDSS are the decision support systems can help to manage overloaded data and turn them into knowledge, reduce the complexity of the work such as automatic complex workflows, and help to identify obese children while reducing the errors, time, and variety of practices [15].

Continuous usage of the information systems result to the size of the database increasing. Therefore the usage of knowledge discovery and data mining in the database (KDD) for the growing databases is important. KDD attempts to gather knowledge by identifying relations from the data sets to help predictions [15]. KDD utilization is increasing in medical informatics and researchers have used it in many areas such as statistics, machine learning, intelligent databases, data visualization, pattern recognition, and high performance computing [17, 18].

The data mining has been used in the medical domain for other purpose like to

improve the decision making such as diagnostic and prognostic problems in oncology, liver pathology, Neuropsychology, and Gynaecology [19]. For better data analysis and decision support, data mining and decision support can be integrated [20]. The task of detecting associations between risk factors and outcomes in the medical area is a difficult work even for experienced biomedical researcher or health care manager [6]. Data mining usage has helped clinicians to improve their health service by assisting in detecting regularities, trends, and unexpected events from the data [16].The usage of data mining tools with advanced algorithms are popular for pattern discovery in biological data [3]. The biological problems include protein interactions, sequence and gene expression data analysis, drug discovery, discovering homologous sequences or structure, construction of phylogenetic trees, gene finding, gene mapping, and sequence alignment [3].

Machine learning was not fully accepted in the medical community because medical practitioners feel that their work is more complicated using such tools [22]. For an example, different models used in healthcare applications have a different explanation especially for model-specific methods [22]. Therefore, important things that must be considered when developing an application for medical practitioners are simplicity and the way of explaining the decisions. Simple techniques used for medical predictions have shown reasonable results [23]. Another challenge is that the systems must able to present discovered knowledge in an easy and fast manner [16].

The data that can be captured by a patient record are classified in three groups: a structured data, semi-structured data, and unstructured data (Figure 6) [22]. Data mining and knowledge discovery techniques and tools based on rule induction are important to analyze the growing size of clinical data.

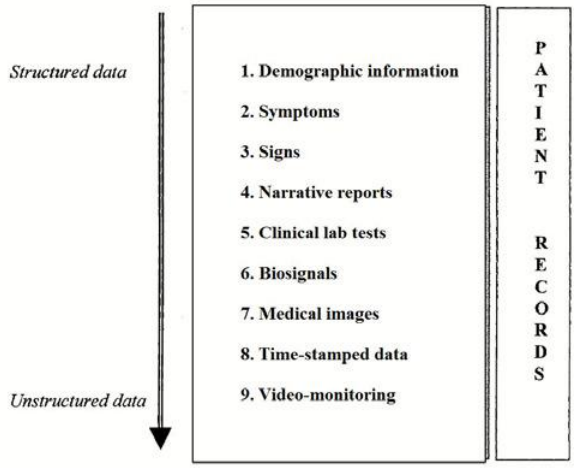


Fig.5: Data That Can be Captured From a Patient Record [16].

5 Well-performing DM algorithms used for disease diagnosisand prognosis

The graphs in Figures 6 and 7 show the most well-performing algorithms used fordisesase diagnosis and prognosis respectively.Diseases in Heart Diseases are classified (Cardiovascular disease, Heart Attack, Coronary Arthery Disease, Hypertension)[7][49], Cancer Diseases (Breast, Prostate, Pancreatic Cancer) and Other Diseases (Asthma, Diabetes, Hepatitis, Kidney Disease, Nerve Diseases, Chronic Disease, Skin Diseases)[3].

As we can see in Fig.4, ANNs are the most well-performing in diagnosing CancerDiseases, Bayesian Algorithms and Decision Trees in Heart Diseases, and DTS indiagnosing other diseases.

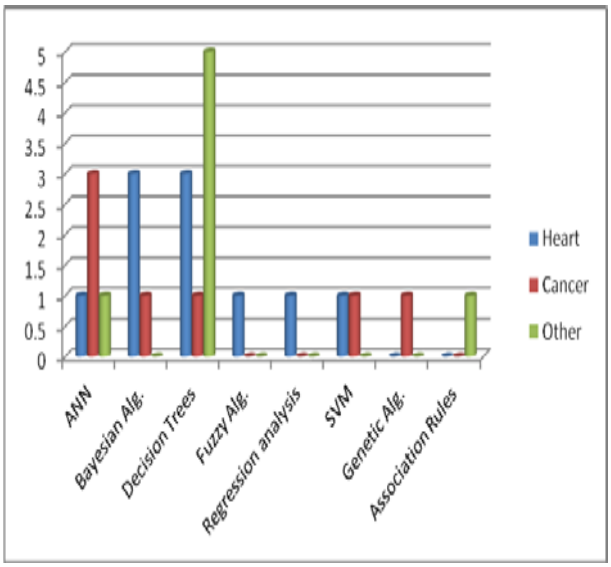
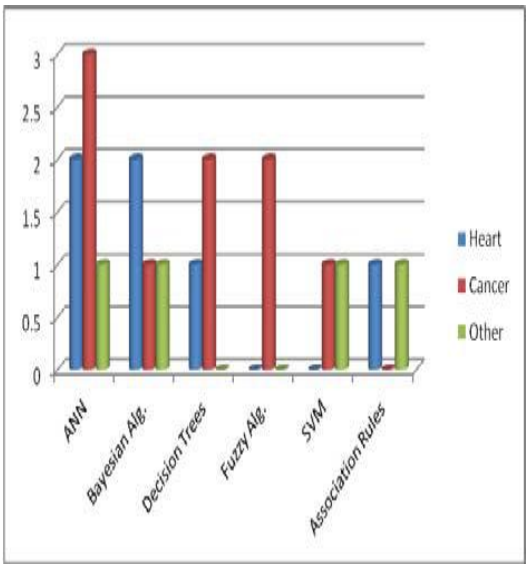


Fig.6 Efficient Algorithms for Disease Diagnosis[3]

On the other side in Fig. 5 it can be seen that for Cancer and Heart Disease Prognosis[52], ANNs are the most well-performing and also Bayesian Algorithms the most well-performing in Heart Diseases Prognosis.



.Fig.7Efficient Algorithms for Disease Prognosis[3]

6. Conclusions

In this paper it is identified and evaluated that the most commonly used DM algorithmsresulting as well-performing on medical databases, based on recent studies. The following algorithms have been identified: Decision Trees (DT's), Artificial neural networks (ANNs) and their Multilayer Perceptron model, Bayesian Networks and Naïve Bayes. Analyses show that DTs, ANNs and Bayesian Algorithms are the most wellperforming algorithms used for disease diagnosis, while ANNs are also the most wellperforming algorithms used for disease prognosis, followed by Bayesian Algorithms,DTs and Fuzzy Algorithms. But it is very difficult to name a single DM algorithm asthe best for the diagnosis and/or prognosis of all diseases. Depending on concrete situations, sometime some algorithms perform better than others, but there are cases when a combination of the best properties of some of the previously mentioned algorithms results more effective.

7. References

[1] Fayyad, U. M. ,Piatetsky-Shapiro, G., Smyth, P., Uthurusamy , R. G. R.: Advances inKnowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA.(1996)

[2]ShantakumarB.Patil, Y.S.Kumaraswamy: Intelligent and EffectiveHeart Attack PredictionSystem Using Data Mining and Artificial Neural Network, European Journal of ScientificResearch ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing,Inc. 2009.

[3] Elma Kolçe (Çela) ,NekiFrasheri:A Literature Review of Data Mining Techniques Used in Healthcare Databases,ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857-7288

[4] Chaitrali S. DangareSulabha S. Apte "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[5] Dimple, Dr. Rahul Rishi:Heart Disease Prediction System Using Multilayer Perceptron, International Journal in Multidisciplinary and Academic Research (SSIJMAR) Vol. 2, No. 4, July- August 2013 (ISSN 2278 – 5973)

[6] Dimple, Dr. Rahul Rishi: Classification of Data for Heart Disease Prediction System Using MLP, International Journal in Multidisciplinary and Academic Research (SSIJMAR) Vol. 2, No. 4, Jan-Feb 2013 (ISSN 2278 – 5973)

[7] M.Kumari, S. Godara: Comparative Study of Data Mining Classification Methods inCardiovascular Disease Prediction, IJCST ISSN : 2229- 4333 Vol. 2, Issue 2, June 2011

[8]K.Srinivas , B.Kavihta Rani, Dr. A.Govrdhan: Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02,(2010),pp 250-255

[9]NidhiBhatla, KiranJyoti," An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 1 Issue 8, October – 2012

[10]J.Soni, U. Ansari, D. Sharma, S. Soni: Predictive Data Mining for Medical Diagnosis: AnOverview of Heart Disease Prediction (2011)

[11]K.S.Kavitha ,K.V.Ramakrishnan , M. K. Singh: Modeling and design of evolutionary neuralnetwork for heart disease detection, IJCSI International Journal of Computer ScienceIssues, Vol. 7, Issue 5, September 2010, ISSN (Online): 1694-0814, pp. 272-283 (2010)

[12]A.A. Aljumah, M. G.Ahamad, M.K.Siddiqui: Predictive Analysis on Hypertension TreatmentUsing Data Mining Approach in Saudi Arabia, Intelligent Information Management,3, (2011), pp. 252-261

[13] [13] J. Chen, et al. (2007). A comparison of four data mining models: bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. 4491/2007.

[14] I. Maglogiannis, et al., "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers,"Applied intelligence, vol. 30, 2007.

[15]MuhamadHarizMuhamadAdnan,Wahidah Husain, Nur'Aini Abdul Rashid," Data Mining for Medical Systems: A Review" Proc. of the International Conference on Advances in Computer and Information Technology - ACIT 2012

[16] A. S. Elmaghraby, et al. (2006). Data Mining from multimedia patient records. 6.

[17]J. C. Prather, et al., "Medical data mining: knowledge discovery in aclinical data warehouse," in AMIA Annual Fall Symposium 1997, pp. 101-105.

[18] J. Han and M. Kamber, Data Mining, concepts and techniques, 1st ed.: Academic Press, 2001.

[19] Yue Huang, et al., "Evaluation of outcome prediction for a clinical diabetes database ", ed, 2004.
[20] A. Pur, et al., "Data mining for decision support: an application in public health care," 2005.
[21] L. Jiang, et al., "A novel bayes model: hidden naive bayes," IEEE Trans. on Knowl. and Data Eng., vol. 21, pp. 1361-1371, 2009.

[22] E. Strumbelj, et al., "Explanation and reliability of prediction models: the case of breast cancer recurrence," Knowl. Inf. Syst., vol. 24, pp. 305-324, 2010.

[23] D. Gregori, et al., "Using Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better?," Journal of Medical Systems, 2011.

[24] 24. N.Satyanandam, Dr. Ch. Satyanarayana, Md.Riyazuddin, A.Shaik: Data Mining Machine Learning Approaches and Medical Diagnose Systems : A Survey

[25] F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O.: Challenges in Data Mining on Medical Databases IGI Global pp. 502-511(2009)

[26] D.Delen: Analysis of cancer data: a data mining approach (2009)

[27] E.Dincer, N.Duru: Prototype of a tool for analysing laryngeal cancer operations

[28] Acute Coronary Syndrome Prediction Using Data Mining Techniques- An Application, World Academy of Science, Engineering and Technology 59 pp.474-478 (2009)

[29] A.O. Osofisan ,O.O. Adeyemo, B.A. Sawyerr, O. Eweje: Prediction of Kidney Failure Using Artificial Neural Networks (2011)

[30] R. Parvathi, S. Palaniammalì: An Improved Medical Diagnosing Technique Using Spatial Association Rules, European Journal of Scientific Research ISSN 1450-216X Vol.61 No.1 pp. 49-59 (2011)

[31]F.I.Dakheel, R.Smko, K. Negrat, A.Almarimi: Using Data Mining Techniques for Finding Cardiac Outlier Patients (2011)

[32] S.K. Wasan, V. Bhatnagar ,H.Kaur: The Impact Of Data Mining Techniques On Medical Diagnostics, Data Science Journal, Volume 5, pp. 119-126 (2006)

[33]S.Palaniappan, R. Awang: Intelligent Heart Disease Prediction System Using Data Mining Techniques (2008)

[34] M.G. Tsipouras, T.P. Exarchos, D.I. Fotiadis,A.P. Kotsia, K.V. Vakalis, K.K. Naka, L. K.Michalis: Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling (2008)

[35] M. L.Jimenez , J. M. Santamarı, R. Barchino, L. Laita, L.M. Laita, L. A. Gonza´lez, A. Asenjo: Knowledge representation for diagnosis of care problems through an expert system: Model of the auto-care deficit situations, Expert Systems with Applications 34 pp.2847–2857 (2008)

[36] M.-J. Huang, M.-Y.Chen, S.-C. Lee: Integrating data mining with case-based reasoning for chronicdiseases prognosis and

diagnosis, Expert Systems with Applications 32 pp.856–867 (2007)

[37]K.Aftarczuk: Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems (2007).

[38]T.Sakthimurugan, S.Poonkuzhali: An Effective Retrieval of Medical Records using Data Mining Techniques, International Journal Of Pharmaceutical Science And Health Care. ISSN: 2249-5738. 2(2), pp 72-78 (2012)

[39]J.Gao, J. Denzinger, and R.C. James: A Cooperative Multi-agent Data Mining Model and Its Application to Medical Data on Diabetes (2005)

[40]A.Habrard, M.Bernard, F. Jacquenet: Multi-Relational Data Mining in Medical Databases, Springer-Verlag (2003), LNAI 278 ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857-7288 581

[41]A.Kusiak, Decomposition in Data Mining: A Medical Case Study , B.V. Dasarathy (Ed.), Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, Vol. 4384, SPIE, Orlando, FL, April 2001, pp. 267-277

[42]S.Floyd: Data Mining Techniques for Prognosis in Pancreatic Cancer (2007)

[43]A.Kika, B.Cico, R.Alimehmeti: Using Machine Learning for Preoperative Peripheral Nerve Surgical Prediction (2010)

[44] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, "Knowledge Management, Data Mining, and Text Mining In Medical Informatics", Chapter 1, eds. Medical Informatics: Knowledge Management And Data Mining In Biomedicine, New York, Springer, pp. 3-34, 2005.

[45] Chaitrali S. DangareSulabha S. Apte “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques” International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[46]Q. Luo, "Advancing knowledge discovery and data mining," in WKDD '08 Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, Washington, DC, USA, 2008.

[47] E. Papagergiou, et al., "Data mining: a new technique in medical research," International Journal of Endocrinology and Metabolism, pp. 189-191, 2005.

[48] A. K. Jain, et al. Artificial neural network : a tutorial [Online].

[49] Y. Xing, et al., "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," presented at the Proceedings of the 2007 International Conference on Convergence Information Technology, 2007.

[50] W. Peng, et al. An Implementation of ID3 --- decision tree learning algorithm [Online].

[51] L. Jiang, et al., "A novel bayes model: hidden naive bayes," IEEE Trans. on Knowl. and Data Eng., vol. 21, pp. 1361-1371, 2009.

[52] J. Chen, et al. (2007). A comparison of four data mining models: bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. 4491/2007.

[53] J. C. Prather, et al., "Medical data mining: knowledge discovery in a clinical data warehouse," in AMIA Annual Fall Symposium 1997, pp. 101-105.