



Methodologies of Data Mining in Healthcare: A Inspection

Dr. A. R. Pon Periasamy

Associate Professor of Computer Science
Nehru Memorial College
Puthanampatti, Trichy (DT)
Tamilnadu, India

S. Mohan

Assistant Professor of Computer Science
Dr.S.Ramadoss Arts & Science College
Periyavadavadi, Vridhachalam
Tamilnadu, India

Abstract— *Data mining is gaining popularity in disparate research fields due to its boundless applications and approaches to mine the data in an appropriate manner. Owing to the changes, the current world acquiring, it is one of the optimal approach for approximating the nearby future consequences. Along with advanced researches in healthcare monstrous of data are available, but the main difficulty is how to cultivate the existing information into a useful practices. To unfold this hurdle the concept of data mining is the best suited. Data mining have a great potential to enable healthcare systems to use data more efficiently and effectively. Hence, it improves care and reduces costs. This paper reviews various Data Mining techniques such as classification, clustering, association, regression in health domain. It also highlights applications, challenges and future work of Data Mining in healthcare.*

Key Words: *Data Mining, Classification, Clustering, Association and Healthcare*

I. INTRODUCTION

In the early 1970's, it was very costly to store the data or information. But due to the advancement in the field of information gathering tools and WWW in the last twenty-five years, we have seen huge amount of information or data are available in electronic format. To store such a large amount of data or information the sizes of databases are increased very rapidly. Such type of databases consist very useful information. This information may be very useful for decision making process in any field. It becomes possible with the help of data mining or Knowledge Discovery in Databases (KDD). Data mining is the process of extracting the useful information from a large collection of data which was previously unknown [1]. A number of relationships are hidden among such a large collection of data for example a relationship between patient data and their number of days of stay [2]. With the help of figure 1 five stages are identified in knowledge discovery process [3, 4, and 5]. With the help of raw data the first stage starts and ends with extracted knowledge which was captured as a result of following stages as shown in figure 1:

Selection

The data is selected according to some criteria in this stage. For example, a bicycle owns by all those people, we can determine subsets of data in this way.

Preprocessing

This stage removes that information which is not necessary for example while doing pregnancy test it is not necessary to note the sex of a patient. It is also known as data cleansing stage.

Transformation

This stage transformed only those data which are useful in a particular research for example only data related to a particular demography is useful in market research.

Data mining

Data mining is a stage knowledge discovery process. This stage is useful for extracting the meaningful patterns from data.

Interpretation and evaluation

The meaningful patterns which the system identified are interpreted into knowledge in this stage. This knowledge may be then useful for making useful decisions.

1.1 Essential of Data Mining in Healthcare

Generally all the healthcare organizations across the world stored the healthcare data in electronic format. Healthcare data mainly contains all the information regarding patients as well as the parties involved in healthcare industries. The storage of such type of data is increased at a very rapidly rate. Due to continuous increasing the size of

electronic healthcare data a type of complexity is exist in it. In other words, we can say that healthcare data becomes very complex. By using the traditional methods it becomes very difficult in order to extract the meaningful information from it. But due to advancement in field of statistics, mathematics and very other disciplines it is now possible to extract the meaningful patterns from it. Data mining is beneficial in such a situation where large collections of healthcare data are available.

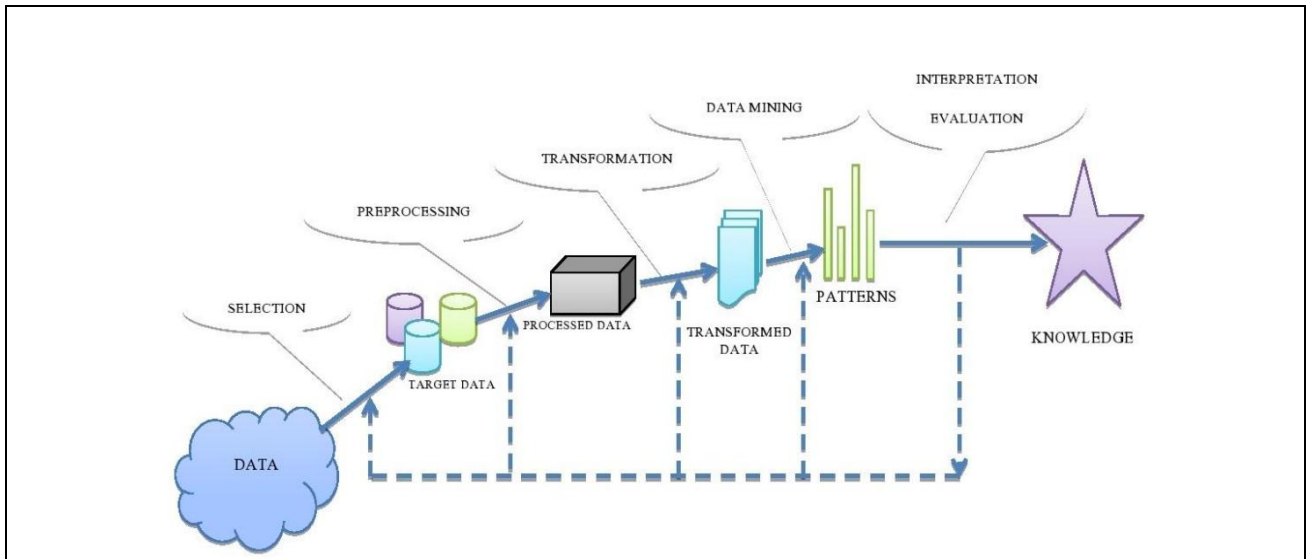


Figure 1: Stages of Knowledge Discovery Process

Data Mining mainly extracts the meaningful patterns which were previously not known. These patterns can be then integrated into the knowledge and with the help of this knowledge essential decisions can be made possible. A number of benefits are provided by the data mining. Some of them are as follows: it plays a very important role in the detection of fraud and abuse, provides better medical treatments at reasonable price, detection of diseases at early stages, intelligent healthcare decision support systems etc. Data mining techniques are very useful in healthcare domain. They provide better medical services to the patients and helps to the healthcare organizations in various medical management decisions. Some of the services provided by the data mining techniques in healthcare are: number of days of stay in a hospital, ranking of hospitals, better effective treatments, fraud insurance claims by patients as well as by providers, readmission of patients, identifies better treatments methods for a particular group of patients, construction of effective drug recommendation systems, etc [2]. Due to all these reasons researchers are greatly influenced by the capabilities of data mining. In the healthcare field researchers widely used the data mining techniques. There are various techniques of data mining. Some of them are classification, clustering, regression, etc. Each and every medical information related to patient as well as to healthcare organizations is useful. With the help of such a powerful tool known as data mining plays a very important role in healthcare industry. Recently researchers use data mining tools in distributed medical environment in order to provide better medical services to a large proportion of population at a very low cost, better customer relationship management, better management of healthcare resources, etc. It provides meaningful information in the field of healthcare which may be then useful for management to take decisions such as estimation of medical staff, decision regarding health insurance policy, selection of treatments, disease prediction etc., [6-9]. Dealing with the issues and challenges of data mining in healthcare [10, 11]. In order to predict the various diseases effective analysis of data mining is used [12-21]. Proposed a data mining methodology in order to improve the result [22-24] and proposed new data mining methodology [25, 26] and proposed framework in order to improve the healthcare system [27-31].

II. DATA MINING METHODOLOGIES

2.1 Types of Classification

Classification is one of the most popularly used methods of Data Mining in Healthcare sector. It divides data samples into target classes. The classification technique predicts the target class for each data point. With the help of classification approach a risk factor can be associated to patients by analyzing their patterns of diseases. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. It consists of predicting a certain outcome based on a given input. Training set is the algorithm which consists of a set of attributes in order to predict the outcome. In order to predict the outcome it attempts to discover the relationship between attributes. Goal or prediction is its outcome. There is another algorithm known as prediction set. It consists of same set of attributes as that of training set. But in prediction set, prediction attribute is yet to be known. In order to process the prediction it mainly analyses the input. The term which defines how "good" the algorithm is its accuracy. Consider a medical database of Pawti Medical Center, training set consists all the information regarding patient which were recorded previously. Whether a

patient had a heart problem or not is the prediction attribute there. With the help of table 1 given below we demonstrates the training sets of such database.

Table 1 – Training And Prediction Sets For Pawti Medical Database
 Training Set

Age	Heart rate	Blood pressure	Heart problem
45	75	140/64	Yes
28	85	101/60	No
38	62	105/55	No

Prediction Set

Age	Heart rate	Blood pressure	Heart problem
33	89	142/82	?
45	52	102/56	?
87	83	138/61	?

In order to disembosom the knowledge, classification predicts rules. Prediction rules are divulged in the form of IF-THEN rules. With the help of above example, a rule predicting the first row in the training set may be represented as follows:

IF (Age=45 AND Heart rate>75) OR (Age>44 AND Blood pressure>139/60) THEN Heart problem=yes

Following are the various classification algorithms used in healthcare:

2.1.1 K-Nearest Neighbor (K-NN) Classification

K-Nearest Neighbor (K-NN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbor) and classified data points according to the voting system [6]. Consider there are various objects. It would be beneficial for us if we know the characteristics features of one of the objects in order to predict it for its nearest neighbors because nearest neighbor objects have similar characteristics. The majority votes of K-NN can play a very important role in order to classify any new instance, where k is any positive integer (small number). It is one of the most simple data mining techniques. It is mainly known as Memory-based classification because at run time training examples must always be in memory [32]. Euclidean distance is calculated when we take the difference between the attributes in case of continuous attributes. But it suffers from a very serious problem when large values bear down the smaller ones. Continuous attributes must be normalized in order to take over this major problem so that they have same influence on the distance measure between distances [33].

K-NN has a number of applications in different areas such as health datasets, image field, cluster analysis, pattern recognition, online marketing *etc.* There are various advantages of KNN classifiers. These are: ease, efficacy, intuitiveness and competitive classification performance in many domains. If the training data is large then it is effective and it is robust to noisy training data. A main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. If there is a big sample then its response time on a sequential computer will also large.

2.1.2 Decision Tree (DT)

DT is considered to be one of the most popular approaches for representing classifier. We can construct a decision tree by using available data which can deal with the problems related to various research areas. It is equivalent to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. Root node is the top most node of a decision tree. For example, with the help of medical readmission decision tree we can decide whether a particular patient requires readmission or not. Knowledge of domain is not required for building decision regarding any problem. The most common use of Decision Tree is in operations research analysis for calculating conditional probabilities [34]. Using Decision Tree, decision makers can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [35]. Decision Tree is widely used by many researchers in healthcare field. Several advantages of decision tree as follows: Decision trees are self-explanatory and when compacted they are also easy to follow. Even set of rules can also be constructed with the help of decision trees. Hence, representation of decision tree plays a very important role in order to represent any discrete-value classifier because it can be capable to handle both type of attributes, nominal as well as numeric input attributes. If any datasets have missing or erroneous values, such type of datasets can be easily handled by decision trees. Due to this reason decision tree can be considered to be nonparametric method. The meaning of above sentence is that there is no need to make assumptions regarding distribution of space and structure of classifier. Decision trees have several disadvantages. These are as follows: Most of the algorithms (like ID3 and C4.5) require that the target attributes have only discrete values because decision trees use the divide and conquer method. If there are more complex interactions

among attributes exist then performance of decision trees is low. Their performance is better only when there exist a few highly relevant attributes. One of the reasons for this is that other classifiers can compactly describe a classifier that would be very challenging to represent using a decision tree. A simple illustration of this phenomenon is replication problem of decision trees [36], and the greedy characteristic of decision trees leads to another disadvantage. This is its over-sensitivity to the training set, irrelevant attributes and to noise [37].

2.1.3 Support Vector Machine (SVM) Available Algorithms

Vapnik *et al.* [38- 39] was the first one who gave the notion of SVM. Among all the available algorithms it provides very accurate results. It is gaining popularity because it can be easily extended to problems related to multiclass though it was mainly developed for problems related to binary classification [40-41]. In order to be useful for various effective and efficient tasks it is capable of creating single as well as multiple hyper planes in high dimensional space. The main aim of creating hyper plane by SVM in order to separate the data points. There are two ways of implementing SVM. The first technique employs mathematical programming and second technique involves kernel functions. With the help of training datasets non linear functions can be easily mapped to high dimensional space. Such situation can only be possible when we use kernel functions. Gaussian, polynomial, sigmoid etc. are some examples of kernel functions. For the classification of data points hyper plane is used. The primary task of hyper plane is to maximize the separation between data points. Support vectors are used in order to construct the hyper plane. There are various advantages of SVM. Some of them as follow: First one is that, it is effective in high dimensional spaces. Another one is that, it is effective in cases where number of dimensions is greater than the number of samples. Third one is that, it is memory efficient because it uses a subset of training points in the decision function (called support vectors), and next is that, it is versatile because different kernel functions can be specified for the decision function. Some of the disadvantages of SVM as follow: First one is that, if the number of features is much greater than number of samples then SVM is more likely to give poor performances, and the another one is that, it does not directly provide probability estimates. These are calculated using an expansive five –fold cross-validation.

2.1.4 Neural Network (NN) Support (SVM)

In the early 20th century it was developed [42]. Before the introduction of decision trees and the Support Vector Machine (SVM) it was regarded as the best classification algorithm [43]. This was one of the reasons which encouraged NN as the most widely used classification algorithm in various biomedicine and healthcare fields [44, 45, 46]. For example, NN has been widely used as the algorithm supporting the diagnosis of diseases including cancers [47-51] and predict outcomes [52-54]. In NN, basic elements are neurons or nodes. These neurons are interconnected and within the network they worked together in parallel in order to produce the output functions. From existing observations they are capable to produce new observations even in those situations where some neurons or nodes within the network fails or go down due to their capability of working in parallel. An activation number is associated to each neuron and a weight is assigned to each edge within a neural network. In order to perform the tasks of classification and pattern recognition neural network is mainly used [55].

The basic property of an NN is that it can minimize the error by adjusting its weight and by making changes in its structure. It minimizes the error only due to its adaptive nature. NN are capable to produce predictions of greater accuracy. One of the major advantages of NN is that it can properly handle noisy data for training and can reasonably classify new types of data that is different from training data. There are also various disadvantages of NN. First is that it requires many parameters, including the optimum number of hidden layer nodes that are empirically determined, and its classification performance is very sensitive to the parameters selected [56]. Second is that its training or learning process is very slow and computationally very expensive. Another is that they do not provide any internal details regarding to that phenomena which is currently under investigation. Hence, this approach is like a “black-box” approach for us.

2.1.5 Bayesian Methods Statistics Plays Role

For probabilistic learning method Bayesian classification is used. With the help of classification algorithm we can easily obtained it [6]. Bayes theorem of statistics plays a very important role in it. While in medical domain attributes such as patient symptoms and their health state are correlated with each other but Naïve Bayes Classifier assumes that all attributes are independent with each other. This is the major disadvantage with Naïve Bayes Classifier. If attributes are independent with each other then Naïve Bayesian classifier has shown great performance in terms of accuracy. In healthcare field they play very important roles. Hence, researchers across the world used them there are various advantages of BBN. One of them is that it helps to makes computation process very easy. Another one is that for huge datasets it has better speed and accuracy.

2.2 Regression & Correlates

Regression is very important technique of data mining. With the help of it, we can easily identify those functions that are useful in order to demonstrates the correlation among different various variables. It is mainly a mathematical tool. With the help of training dataset we can easily construct it. Consider two variables ‘P’ and ‘Q’. These two types of variables are mainly used in the field of statistics. One of them is known as dependent and another one is independent variables. The maximum number of dependent variables cannot be more than one while independent can be exceeds one. Regression is mostly used in order to inspect the certain relationship between variables. With the help of regression technique we can easily entrenched the addiction of one variable upon others [35]. Regression can be classified into linear and non-linear on the basis of certain count of independent variables. In order to appraisal associations between two types of variables in which one is dependent variable and another one is independent variables

(one or more), linear regression used. In order to construct the linear model, linear function is utilized by linear regression. But there is limitation while we use linear approach because both types of variables are known already and hence, its main purpose is to trace a line that correlates between both these variables [46]. We cannot use linear regression for categorized data. It is restricted only to numerical data. With the help of logistic regression the categorical data can be used. Such type of data is used by non-linear regression and logistic regression is basically a type of non-linear regression. Logistic regression with the help of logit function can predict the probability of occurrence. However, between variables logistic regression cannot consider linear relationship [47]. Due to all these reasons regression is widely used in medical field for predicting the diseases or survivability of a patient.

Regression Illution in Healthcare

Divya *et al.*, proposed Weighted Support Vector Regression (WSVR) which used weight factor on the basis of sensor reading for providing continuous monitoring to patients in order to provide them better healthcare services. In this research, on the basis of experimental result, it has been found that the proposed approach had better accuracy than simple vector regression [58].

Logistic regression for the estimation of relative risk for various medical conditions such as Diabetes, Angina, stroke etc [59].

Xie *et al.*, proposed a regression decision tree algorithm in order to predict the number of hospitalization days in a population. Proposed algorithm was developed using extensive health insurance claims data sets. Experimental results displayed that proposed algorithm was performed significantly in general population as well as in sub-populations in order to predict future hospitalization [60].

Alapont *et al.*, used WEKA for different learning methods like LinearRegression, LeastMedSq, SMOreg, Multilayer Perceptron, KStart, Tree M5P etc. In the research, experiments were carried out using 10-fold cross validation. After experiments were carried out, it has been found that LinearRegression and Tree M5P gave best results for the effective utilization of hospital resources, improved hospital ranking and better customer relationship services [61].

2.3 Clustering Dependent & Independent

Clustering is defined as unsupervised learning that occurs by observing only independent variables while supervised learning analyzing both independent and dependent variables. It is different from classification which is a supervised learning method. It has no predefined classes. Because of this reason, clustering may be best used for studies of an exploratory nature, mainly if those studies encompass large amount of data, but not very much known about data (such as mass of data generated by microarray analysis). The goal of clustering is descriptive while goal of classification is predictive (Veyssieres and Plant, 1998). The main task of unsupervised learning method means clustering method is to form the clusters from large database on the basis of similarity measure [6]. The goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, an important part of the assessment is extrinsic. Clustering partitioned the data points based on the similarity measure [6]. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belongs to different groups. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it grasp various scientific disciplines: from mathematics and statistics to biology and genetics, each of which uses different terms to describe the topologies formed using this analysis. From biological “taxonomies”, to medical “syndromes” and genetic “genotypes” to manufacturing “group technology”— the problem is identical: forming categories of entities and assigning individuals to the proper groups within it. Following are the various clustering algorithms used in healthcare.

III. DATA MINING FEATURES IN HEALTHCARE

As we know that a lot of healthcare data is generated and stored by various healthcare organizations. But there are various challenges related to healthcare data which may play serious hurdles in the making proper decisions. The first challenge with healthcare data is the format of data being stored is different in different healthcare organizations. Till date there is no standard format is laid down for data being stored. In epidemic situations this lack of standard format can make the epidemic situations even more worse. Suppose that an epidemic disease is spread within a country at its different geographical regions. The country health ministry requires that all the healthcare organizations must share their healthcare data with its centralized data warehouse for analysis in order to take all the essential steps so that epidemic situation may get resolve. But since the formats of data is different. Hence, the analysis of data may take longer time than usual. Due to this it may be possible that the situation may become out of control. The healthcare data is very useful in order to extract the meaningful information from it for improving the healthcare services for the patients. To do this quality of data is very important because we cannot extract the meaningful information from that data which have no quality. Hence, the quality of data is another very important challenge. The quality of data depends on various factors such as removal of noisy data, free from missing of data etc.

All the necessary steps must be taken in order to maintain the quality in healthcare data. Data sharing is another major challenge. Neither patients nor healthcare organizations are interested in sharing of their private data. Due to this the epidemic situations may get worse, planning to provide better treatments for a large population may not be possible, and difficulty in the detection of fraud and abuse in healthcare insurance companies etc. Another challenge is that in order to build the data warehouse where all the healthcare organizations within a country share their data is very costly and time consuming process.

IV. CONCLUSION AND FUTURE WORK

For any algorithm its accuracy and performance is of greater importance. But due to presence of some factors any algorithm can greatly lost the above mentioned property of accuracy and performance. Classification is also belongs to such an algorithm. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm but also degrades its performance. Hence, before applying classification algorithm it must be necessary to remove all those attributes from datasets who later on acts as noisy attributes. Feature selection methods play a very important role in order to select those attributes who improves the performance of classification algorithm.

Clustering techniques are very useful especially in pattern recognitions. But they suffer from a problem on choosing the appropriate algorithm because regarding datasets they do not have information. We can choose partitionial algorithm only when we know the number of clusters. Hierarchical clustering is used even when we do not know about the number of clusters. Hierarchical clustering provides better performance when there is less datasets but as soon as volume of datasets increases its performance degrades. To overcome this problem random sampling is very beneficial.

In hierarchical clustering, if the data is too large to be presented in a dendrogram, the visualization capability is very poor. One possible solution to this problem is to randomly sample the data so that users can properly understand the overall grouping/similarity of the data using the dendrogram that is generated with the sampled data. The main drawback to the use of hierarchical clustering algorithms is cubic time complexity. This complexity is such that the algorithms are very much limited for very large data sets. As the result, the hierarchical algorithms are much slower (in computational time) than partitionial clustering algorithms. They also use a huge amount of system memory to calculate distances between objects.

The privacy regarding to patient's confidential information is very important. Such type of privacy may be lost during sharing of data in distributed healthcare environment. Necessary steps must be taken in order to provide proper security so that their confidential information must not be accessed by any unauthorized organizations. But in situations like epidemic, planning better healthcare services for a very large population etc. some confidential data may be provided to the researchers and government organizations or any authorized organizations.

In order to achieve better accuracy in the prediction of diseases, improving survivability rate regarding serious death related problems etc. various data mining techniques must be used in combination.

To achieve medical data of higher quality all the necessary steps must be taken in order to build the better medical information systems which provides accurate information regarding to patients medical history rather than the information regarding to their billing invoices. Because high quality healthcare data is useful for providing better medical services only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry.

Takes all necessary steps in order to minimize the semantic gap in data sharing between distributed healthcare databases environment so that meaningful patterns can be obtained. These patterns can be very useful in order to improve the treatment effectiveness services, to better detection of fraud and abuse, improved customer relationship management across the world.

REFERENCES

- [1] D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, (2001).
- [2] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005). [3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD process of extracting useful knowledge form volumes of data.commun.", ACM, vol. 39, no. 11, (1996), pp. 27-34.
- [4] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, (2006).
- [5] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases", Commun. ACM, vol. 39, no. 11, (1996), pp. 24-26.
- [6] C. McGregor, C. Christina and J. Andrew, "A process mining driven framework for clinical guideline improvement in critical care", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). <http://eur-ws.org>, vol. 765, (2012).
- [7] P. R. Harper, "A review and comparison of classification algorithms for medical decision making", Health Policy, vol. 71, (2005), pp. 315-331.
- [8] V. S. Stel, S. M. Pluijm, D. J. Deeg, J. H. Smit, L. M. Bouter and P. Lips, "A classification tree for predicting recurrent falling in community-dwelling older persons", J. Am. Geriatr. Soc., vol. 51, (2003), pp. 1356-1364.
- [9] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines", Int. J. Med. Inform., vol. 77, (2008), pp. 81-97.
- [10] R. D. Canlas Jr., "Data Mining in Healthcare:Current Applications and Issues", (2009).
- [11] F. Hosseinkhah, H. Ashktorab, R. Veen, M. M. Owrang O., "Challenges in Data Mining on Medical Databases", IGI Global, (2009), pp. 502-511.
- [12] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST ISSN: 2229- 4333, vol. 2, no. 2, (2011) June.

- [13] J. Soni, U. Ansari, D. Sharma and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", (2011).
- [14] C. S. Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques", (2012).
- [15] K. Srinivas, B. Kavihta Rani and Dr. A.Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering, vol. 02, no. 02, (2010), pp. 250-255.
- [16] A. Aljumah, M. G.Ahamad and M. K. Siddiqui, "Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia", Intelligent Information Management, vol. 3, (2011), pp. 252-261.
- [17] D. Delen, "Analysis of cancer data: a data mining approach", (2009).
- [18] O. Osofisan, O. O. Adeyemo, B. A. Sawyerr and O. Eweje, "Prediction of Kidney Failure Using Artificial Neural Networks", (2011).
- [19] S. Floyd, "Data Mining Techniques for Prognosis in Pancreatic Cancer", (2007).
- [20] M.-J. Huang, M.-Y. Chen and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", Expert Systems with Applications, vol. 32, (2007), pp. 856-867.
- [21] S. Gupta, D. Kumar and A. Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis", (2011).
- [22] K. S. Kavitha, K. V. Ramakrishnan and M. K. Singh, "Modeling and design of evolutionary neural network for heart disease detection", IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, vol. 7, no. 5, (2010) September, pp. 272-283.
- [23] S. H. Ha and S. H. Joo, "A Hybrid Data Mining Method for the Medical Classification of Chest Pain", International Journal of Computer and Information Engineering, vol. 4, no. 1, (2010), pp. 33-38.
- [24] R. Parvathi and S. Palaniammali, "An Improved Medical Diagnosing Technique Using Spatial Association Rules", European Journal of Scientific Research ISSN 1450- 216X, vol. 61, no. 1, (2011), pp. 49-59.
- [25] S. Chao and F. Wong, "An Incremental Decision Tree Learning Methodology Regarding Attributes in Medical Data Mining", (2009).
- [26] Habrard, M. Bernard and F. Jacquenet, "Multi-Relational Data Mining in Medical Databases", Springer-Verlag, (2003).
- [27] S. B. Patil and Y. S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X, © EuroJournals Publishing, Inc., vol. 31, no. 4, (2009), pp. 642-656.
- [28] A.Shukla, R. Tiwari, P. Kaur, Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference, IACC 2009.
- [29] L. Duan, W. N. Street & E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterprise Information Systems, 5:2, pp169-181 , 2011.
- [30] D. S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", (2011).
- [31] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (2008).
- [32] Alpaydin, E. (1997), Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, p. 115–132.
- [33] Bramer, M., (2007) Principles of data mining: Springer. [34] Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, <http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM- Classification.pdf>, (2003).
- [35] Apte & S.M. Weiss, Data Mining with Decision Trees and Decision Rules, T.J. Watson Research Center, http://www.research.ibm.com/dar/papers/pdf/fgcsaptewe issue_with_cover.pdf, (1997).
- [36] Pagallo, G. and Huassler, D., Boolean feature discovery in empirical learning, Machine Learning, 5(1): 71-99, 1990.
- [37] Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, 1993.
- [38] V. Vapnik, "Statistical Learning Theory", Wiley, (1998). [39] V. Vapnik, "The support vector method of function estimation", (1998).
- [40] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines, and other kernel-based learning methods", Cambridge University Press, (2000).
- [41] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines", Cambridge University Press, (2000).
- [42] Anderson, J. A., and Davis, J., An introduction to neural networks. MIT, Cambride, 1995.
- [43] Obenshain, M. K., Application of data mining techniques to healthcare data. Infect. Control Hosp. Epidemiol. 25(8):690–695, 2004.
- [44] Bellazzi, R., and Zupan, B., Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inform. 77:81–97, 2008.
- [45] Übeyli, E. D., Comparison of different classification algorithms in clinical decision making. Expert syst 24(1):17–31, 2007.

- [46] Kaur, H., and Wasan, S. K., Empirical study on applications of data mining techniques in healthcare. *J. Comput. Sci.* 2(2):194–200 2006.
- [47] Romeo, M., Burden, F., Quinn, M., Wood, B., and McNaughton, D., Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. *Cell. Mol. Biol. (Noisy-le-Grand, France)* 44(1):179, 1998.
- [48] Ball, G., Mian, S., Holding, F., Allibone, R., Lowe, J., Ali, S., et al., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 18(3):395–404, 2002.
- [49] Aleynikov, S., and Micheli-Tzanakou, E., Classification of retinal damage by a neural network based system. *J. Med. Syst.* 22(3):129–136, 1998.
- [50] Potter, R., Comparison of classification algorithms applied to breast cancer diagnosis and prognosis, advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, July 2007, pp.40–49.
- [51] Kononenko, I., Bratko, I., and Kukar, M., Application of machine learning to medical diagnosis. *Machine Learning and Data Mining: Methods and Applications* 389:408, 1997.
- [52] Sharma, A., and Roy, R. J., Design of a recognition system to predict movement during anesthesia. *IEEE Trans. Biomed. Eng.* 44(6):505–511, 1997.
- [53] Einstein, A. J., Wu, H. S., Sanchez, M., and Gil, J., Fractal characterization of chromatin appearance for diagnosis in breast cytology. *J. Pathol.* 185(4):366–381, 1998. [54] Brickley, M., Shepherd, J. P., and Armstrong, R. A., Neural networks: a new technique for development of decision support systems in dentistry. *J. Dent.* 26(4):305–309, 1998.
- [55] M. H. Dunham, “Data mining introductory and advanced topics”, Upper Saddle River, NJ: Pearson Education, Inc., (2003).
- [56] Schwarzer, G., Vach, W., and Schumacher, M., On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat. Med.* 19:541–561, 2000.
- [57] H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, “A Comparative Study of Classification Methods For Microarray Data Analysis”, Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, (2006), pp. 33-37.
- [58] R. Potter, “Comparison of classification algorithms applied to breast cancer diagnosis and prognosis”, advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, (2007) July, pp. 40-49.
- [59] L. Huang, H. C. Liao and M. C. Chen, “Prediction model building and feature selection with support vector machines in breast cancer diagnosis”, *Expert Systems with Applications*, vol. 34, (2008), pp. 578-587.
- [60] G. Beller, “The rising cost of health care in the United States: is it making the United States globally noncompetitive?”, *J. Nucl. Cardiol.*, vol. 15, no. 4, (2008), pp. 481-482.
- [61] M. U. Khan, J. P. Choi, H. Shin and M. Kim, “Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare”, 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, (2008) August 20-24.
- [62] L. Chang and C. H. Chen, “Applying decision tree and neural network to increase quality of dermatologic diagnosis”, *Expert Systems with Applications*, Elsevier, vol. 36, (2009), pp. 4035-4041.
- [63] R. Das, I. Turkoglu and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles”, *Expert Systems with Applications*, vol. 36, (2009), pp. 7675-7680.
- [64] I. Curiac, G. Vasile, O. Baniac, C. Volosencu and A. Albu, “Bayesian Network Model for Diagnosis of Psychiatric Diseases”, Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, (2009) June 22-25.
- [65] K. F. R. Liu and C. F. Lu, “BBN-Based Decision Support for Health Risk Analysis”, Fifth International Joint Conference on INC, IMS and IDC, (2009).
- [66] Avci, “A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier”, *Expert Systems with Applications*, Elsevier, vol. 36, (2009), pp. 10618-10626.
- [67] S. Gunasundari and S. Baskar, “Application of Artificial Neural Network in identification of Lung Diseases”, *Nature & Biologically Inspired Computing*, 2009. NaBIC 2009. World Congress on. IEEE, (2009).
- [68] S. Soni and O. P. Vyas, “Using Associative Classifiers for Predictive Analysis in Health Care Data.