**Analysing accuracy and performance of data mining techniques in healthcare**

|  |
| --- |
| Vikas Kapoor, M-tech |
| Dept. Of Computer Science |
| ITM University |
| Gurgaon, Haryana |

**ABSTRACT**

Data mining is growing technique in disparate research fields due to its limitless approaches and applications to mining the data in the targeted manner. Owing to the changes, the current world acquiring, it is one of the optimize technique (approach) for predicting future result and its analysis. With advanced researches in healthcare sector monstrous of data are available, but the main difficulty is how to cultivate the existing information into a useful practices. To solve this barrier the concept of data mining is introduced and it is the best technique suited. Data mining have a great potential to enable healthcare systems more efficiently and effectively by using its data. Hence, it improves care of patient and reduces costs involved. This paper analysis various Data Mining techniques such as classification, association, healthcare, clustering and regression . The paper also highlights applications, challenges and future work of Data Mining in healthcare.

**Keywords**

Data Mining, Classification, Clustering, Association, Healthcare, Knowledge discovery

**1. INTRODUCTION**

In the early 1975’s, the storage of data or information was very much. But Because of the advancement in the field of information gathering tools or techniques and priceless work done by WWW in the last twenty-five years, data now can easily be stored electronically and extracted information in the form of knowledge become easy because of data mining techniques.

A database is required to store a large amount of data, which is continuously and rapidly increasing. This information will be very useful in decision making process in any field. Data mining is the process of extracting the useful information from a extremely large collection of data which was previously unknown [1]. A number of relationships are hidden among such collection of data. For example, a relationship between patients and no. of days stayed. [2].

With the help of figure 1 five stages are identified in knowledge discovery process [3, 4, and 5].The data system identifies are interpreted into knowledge in this level or stage.

**Data mining steps:**

First, **Selection** data is selected on the behalf of some criteria in this stage. For example, “a bicycle owns by all those people”, we can determine subsets of data in this way.

Second, in the **preprocessing** stage removes all the information which is actually not necessary in data. This is called data cleansing stage.

Third, in **transformation** stage only those useful data is extracted which is being used in particular research. for example only data related to a particular demography is very useful in market research area.

Fourth, It is a stage **knowledge discovery process**. This stage is useful for extracting the meaningful patterns from large data.

The meaningful patterns analyses is used for making useful decisions.

**1.1 Role of Data Mining in Healthcare**

Usually all the healthcare companies maintain and store their data in electronic formats across the world like Patients and parties involved in healthcare industries. The storage of healthcare type of data is increasing at a very rapidly now-a-days. In other words, we can say that healthcare data becomes very complex and difficult to store. By using the traditional methods it become difficult to extract the meaningful information or knowledge from it. But due to huge advancement in field of data mining of statistics, mathematics and very other disciplines it is now possible to extract the meaningful patterns from it. Data mining is useful to analyze and generate knowledge from healthcare data.

Data Mining normally extracts the useful patterns/knowledge which was previously not known. These patterns can be integrated into the knowledge and with the help of this knowledge , essential decisions can be made now. A lot of benefits are provided by the data mining. Some of them are as follows: it plays a very important role in investigation of detection of fraud and abuse, provides better medical treatments, detection of diseases at earliest possible stages, intelligent healthcare decision support system by using data mining techniques, etc [2]. Due to all these reasons researchers are greatly influenced and invest time and money in Data Mining. In the healthcare field researchers widely used the data mining techniques. Some of them are classification, clustering, regression, etc. Each and every medical information related to patient as well as to healthcare organizations is useful. Now-a-days researchers uses data mining tools in distributed medical environment to provide better medical services to a large proportion of population at a very low cost.It also provides better customer service and relationship management, better and optimize management of healthcare resources.

**2. Techniques of Data Mining**

**2.1 Classification**

Classification is important and popular used methods in Data Mining in the sector of healthcare. It used to divide all input (samples) data into output( target) data classes by using classify dataset. The classification technique predicts the target class for each data points. With the help of classification approach ,a risk factor can be associated to patients by analyzing their patterns of diseases. This is one of the supervised learning approaches with known class category. Multilevel and binary are the two methods of classification. In the binary classification, there are only two possible classes such as, “high classes” or “low classes” while the many class approach has more than two targets. Data set is partitioned as training and testing dataset. It consists of prediction of a certain result in their outcome based on provided input. Training (input) dataset is the algorithm which consists of a set of attributes in order to predict the result (outcome). For the prediction of 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 database.

**Table 1 – TRAINING DATABASE FOR TRAINING MEDICAL DATA**

**Training Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **Heart rate** | **Blood** | **Heart** |
|  |  | **pressure** | **problem** |
|  |  |  |  |
| **45** | **75** | **140/64** | **Yes** |
|  |  |  |  |
| **29** | **83** | **101/60** | **No** |
|  |  |  |  |
| **35** | **69** | **105/57** | **No** |
|  |  |  |  |

**Prediction Set**

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

In order to use the knowledge, classification prediction rules are used. Prediction rules divides the data in the form of IF-THEN rules. With the help of above example, a rule created to predict the first row of database in the training set may be represented as follows:

IF (Age\_of\_person =45 AND Heart\_Rate>75) OR (Age>44 AND Blood pressure of the person >139/60) THEN Heart problem=yes

Else Heart problem=no.

Following are the various classification algorithms used in healthcare:

*2.1.1 K-Nearest Neighbor (KNN)*

(K-NN) K-Nearest Neighbor is the most simplest and easiest classifier that discovers the unidentified data point from the dataset point using the previously known data points (nearest neighbor) and classify data points according to the selection system which is used to classify the dataset. Consider there are various objects. It would be surely 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 positive integer (small number). It is the simplest data mining techniques. It is mainly known as Memory based classification because at run time of training examples,it must be always in memory. 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 .

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, efficiency, 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. The main disadvantage of KNN classifiers is that the large memory is needed to store the whole training sample. It increase the response time of the machine.

*2.1.2* *Decision Tree (DT)*

DT is again the most popular and considered as the best approach 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 or nearby it can be considered as the flowchart in which every non-leaf nodes represent a test on a selected attribute and every branch of leaf denotes an outcome of that test and every leaf node have a class label for representation. The top most root is called root 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 problem. The mainly common use of Decision Tree is in the operations ,research analysis for calculating conditional probabilities. By the use of Decision Tree, decision makers can choose best alternative and traversal from any root to leaf indicates unique class separation based on the maximum information gain. Several advantages of DT are as follows. 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 because it is 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 as 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. DT has several demerits. These are as follows: Most of the algorithms for example (like ID3 and C4.5) require that the target attributes have only the discrete values because DT use the divide and conquer method. If there more complex interactions among attributes exist then performance of decision trees is will be low. Their performance is better only when there exist some of few highly relevant attributes in DT. One of the reasons for this need is that the other classifiers can compactly describe a classifier that would be very much challenging to represent using a decision tree. A simple illustration of this phenomenon is replication problem of decision trees, and the greedy characteristic of decision trees leads to another disadvantage. This is its over-sensitivity to the training set, irreverent attributes and to noise.

*2.1.3 Support Vector Machine (SVM)*

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 is 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 only possible when we use kernel functions for this. 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 very effective in high dimensional spaces. Another one, it is very effective in the cases where number of dimensions is greater than the number of samples input. Third is, it is memory efficient because it uses a subset of training points in the decision function in SVM (called support vectors), and next is that, it has versatile nature because different kernel functions can be specified for the decision function. Some of the disadvantages of SVM are as follows: First one is, if the number of features is much greater than number of samples then SVM is most likely to give poor performances, and second is that, it does not directly provide probability estimates. These are calculated using an expensive fivefold cross validation.

*2.1.4 Neural Network (NN)*

In the late 19th century it was developed [42]. Before the introduction of support vector machine (SVM) and DT (decision trees) it was considered as the best algorithm for classification which has far better result and techniques as compared to previous data This was one of the reasons which automatically encourage NN as the most widely used classification algorithm in various fields like biomedicine,healthcare fields and data classification . For example, NN has been widely used as the algorithm supporting the diagnosis of diseases including cancers, MRI and predicting outcomes [47-52]. In NN, basic elements are neurons or nodes. These neurons are interconnected and within the network they work 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 because of its adaptive nature. NN are capable to produce predictions of greater accuracy. One of best advantages of Neural network (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 which is the best part for neural network. There are also various demerits of Neural network (NN). First is that it requires man yof parameters, including the optimum numbers of hidden layers nodes that are empirically determined, and its classification performance is very sensitive to the parameters selected by NN, Second is that its training or learning process is very much slow and computationally very expensive as compare to SVM.

*2.1.5 Bayesian Methods or Bayesian algorithm*

For Bayesian method ,probabilistic learning method of classification is used. Bayes theorem of statistics is a very important algorithm in the field of data mining and play a very important role in it. 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 of 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 make computation process very easy. Another one is that for huge datasets it has better speed and accuracy.

**2.2 Regression**

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 [85]. 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 [86]. 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 [87].

**Regression Examples in Healthcare**

Divya *et al.,* proposed (WSVR) Weighted Support Vector Regression 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 [88].Xie*,* proposed a regression DT 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 [90].Alapont *et al.,* used WEKA for different learning methods like Linear Regression, LeastMedSq, SMOreg, Multilayer Perception, 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 Linear Regression and Tree M5P gave best results for the effective utilization of hospital resources, improved hospital ranking and better customer relationship services [91].

**2.3 Clustering technique in data mining**

Unsupervised learning that occurs by only observing independent variable while in supervised learning we analyze both the independent and dependent variable .

It is different from classification which is a supervised learning method. It has no predefined classes. The main task of unsupervised learning method means clustering is to form the clusters from large database on the basis of similarity measure. The goal of clustering is to discover a new set of data 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.

*2.3.1 Partitional Clustering*

The maximum number of data points in the datasets is ‘n’. With the help of ‘n’ data points the maximum possible number of ‘k’ clusters is obtained. In order to obtained the ‘k’ clusters from ‘n’ data points partitioned clustering method is used. In this method, each ‘n’ data points is related to one and only ‘k’ clusters while each ‘k’ clusters can relates to more than ‘n’ data points. Partitional clustering algorithms require a user to provide the input input k to the machine, (which is the number of clusters). Normally, partitional algorithms directly relocate objects to k clusters in clustering. Partitional algorithms are categorized according to how they relocate objects in clusters, how they select a cluster centroid (or representative) among objects within a (incomplete) cluster is main point, and how they measure similarities between objects and cluster centroids. Before we obtained the clusters, this method requires to define the required number of cluster which we may have to obtained from datasets. On the basis of similarities between objects and cluster centroids this method is partitioned into two categories. These are K-means and K-Mediods. One of the most popular algorithms of this approach is K-means [3]. First of all it randomly selects the values of k objects and then decomposes these objects values into k disjoint groups by iteratively relocating objects values based on the similarity between the centroids and objects [93]. In k-means, a cluster centriod is mean value of objects in the cluster. The next algorithm is K-mediods. In order to group the cluster ,it uses mediods. Mediod is very important because in the database it is that data point which is most centrally located. In order to improve the healthcare services related to public healthcares domain Lenert *et al.*, utilized the application of k-means clustering [94] and by using the clustering technique Belciug *et al*. detect the recurrence of breast cancer [95]. Escudero *et al.*, used the concept of Bioprofile and K-means clustering forearly detection of Alzheimer’s disease The major advantage of partitioned clustering algorithms is their superior clustering accuracy as compared with hierarchal clustering algorithms that is the result of their global optimization strategy

*2.3.2 Hierarchical Clustering*

In order to partition the data points this method can be used two approaches. Data points can be partitioned in a tree way known as hierarchical way by using either top down or bottom up approaches. On the basis of partitioned process this method can be classified as Agglomerative and Divisive. The maximum number of data points can be n. A number of data points among n data points may have similarity with each other. The main aim of agglomerative approach is to merge such data points into a single group [3]. Divisive approach initially takes this single group and iteratively partitioned it into smaller group until and unless each data point relates to one and only one cluster. There are three types of hierarchical clustering algorithms: First one is single-link hierarchical clustering, another one is complete-link hierarchical clustering, and the last one is average-link. In single-link Hierarchical clustering algorithm select the closet pair of objects from two groups and measure the similarity between objects as group similarity in this clustering. The complete link algorithm is important to calculate the similarity between the most distant pair of objects from two groups. The average-link algorithm selects all pairs of objects from two groups and find averages all possible distances between these objects. Most similar two groups, those having the shortest distance are merged together after calculating similarities or distances between groups. Among the various hierarchical algorithms we find average link algorithm provides the best accuracy in most cases we test. The advantage of the use of hierarchical clustering algorithms is the capability of visualization that shows how much objects in the data set are similar one another. In addition, with the utilization of a dendogram, researchers can reasonably guess the number of clusters.

*2.3.3 Density Based Clustering*

Density based clustering methods play a very important role in biomedical research because they are capable of handing any cluster of arbitrary shape. Recent researches prove that this method can be efficiently and effectively beneficial in order to extract the meaningful patterns from a very large database which mainly consist biomedical images. Besides density based clustering method, partitioned clustering and hierarchical clustering methods do not extract the meaningful patterns from biomedical images database because these two methods are capable to handle only the clusters of spherical shape not the clusters of arbitrary shape. To remove the problem of positional clustering and hierarchical clustering methods ,density based clustering method evolved. On the basis of density distribution function following are the main approaches of density based clustering: DBSCAN, OPTICS, and DENCLUE [3]. Celebi *et al.,* used density based clustering approach in order to obtain the useful patterns from a very large biomedical images database. These patterns play a very important role in order to determine homogeneous colour [99]. There are various advantages of the density based clustering. First one is that in density based clustering approach, in advance the number of clusters are not required. Another one is that, it can easily handle the clusters of any arbitrary shape. And the last one is that it can be used very effectively and efficiently even in noisy situations. In other words it is performed equally well in noisy situations. One major disadvantage of density based clustering is that, if there is a lot of variation in densities along with data points then it cannot be handle such variations in data points. Another disadvantage is that, distance measure is the primary factor which calculates its result.

**3. DATA MINING CHALLENGES 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.

**4. CONCLUSION AND FUTURE WORK**

For above definitions and discussions, in algorithm, its accuracy, working efficiency and performance is of greater importance. But presence of some constraints in every algorithm ,the accuracy and performance of the algorithm is affected. . If there is any noisy data present in dataset then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm of that data but also degrades its performance. Hence, in any case before applying classification algorithm it must be necessary to remove all those attributes from datasets which later on acts as noisy attributes in the datasets. Feature selection methods is also a very important method. It creates a platform for classification. Important role is to select those attributes who improves the performance of classification algorithm.

Clustering techniques are very useful especially in pattern recognitions and classification. But sometimes they suffer from a problem of choosing the right algorithm because regarding datasets they do not have information before. We can choose partitioned 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.

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 has to 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 not only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry.

We should take 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, for better detection of fraud and abuse, improved customer relationship management across the world.

**REFERENCES**

1. Author :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://ceur-ws. org, vol. 765, **(2012)**.

Author: 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.

Author: J. Soni, U. Ansari, D. Sharma and S. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, **(2011)**.

C. S. Dangare and S. S. Apte, “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques”, **(2012)**.

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.

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.

A.Shukla, R. Tiwari, P. Kaur, Knowledge Based Approach for Diagnosis of Breast Cancer, IEEE International Advance Computing Conference,IACC **2009.**

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.

D. S. Kumar, G. Sathyadevi and S. Sivanesh, “Decision Support System for Medical Diagnosis Using Data

Mining”, **(2011)**.

O. Er, N. Yumusakc and F. Temurtas, “Chest diseases diagnosis using artificial neural networks”, Expert Systems with Applications, vol. 37, **(2010)**, pp. 7648-7655.

Chuang, L.Y., Wu, K.C., Chang, H.W. and Yang, C.H. (2011) “Support Vector Machine-Based Prediction for Oral Cancer Using Four SNPs in DNA Repair Genes”. Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong,16-18 March **2011,** 16-18.

A. Bakar, Z. Kefli, S. Abdullah and M. Sahani, “Predictive Models for Dengue Outbreak Using Multiple Rulebase Classifiers”, 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, **(2011)** July 17-19.

S. S. Moon, S. Y. Kang, W. Jitpitaklert and S. B. Kim, “Decision tree models for characterizing smoking patterns of older adults”, Expert Systems with Applications, Elsevier, vol. 39, **(2012)**, pp. 445-451.

H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S. Chen, “Application of classification techniques on development an early-warning systemfor chronic illnesses”, Expert Systems with Applications, vol. 39, **(2012)**, pp. 8852-8858.

Chien and G. J. Pottie, “A Universal Hybrid Decision Tree Classifier Design for Human Activity Classification”, 34th Annual International Conference of the IEEE EMBS San Diego, California USA, **(2012)** August 28-September 1.

M. Shouman, T. Turner and R. Stocker, “Applying K-Nearest Neighbour in Diagnosing Heart Disease Patients”, International Conference on Knowledge Discovery (ICKD-2012), **(2012)**.

Y. Liu, H. L. Chen, B. Yang, X. E. Lv, N. L. Li and J. Liu, “Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease”, Journal of Medical System, Springer, **(2012)**.

Hattice and K. Metin, “A Diagnostic Software tool for Skin Diseases with Basic and Weighted K-NN”, Innovations in Intelligent Systems and Applications (INISTA), **(2012)**.

M. J. Abdi and D. Giveki, “Automatic detection of erythemato-squamous diseases using PSO–SVM based on association rules”, Engineering Applications of Artificial Intelligence, vol. 26, **(2013)**, pp. 603-608.

W. L. Zuoa, Z. Y. Wanga, T. Liua and H. L. Chenc, “Effective detection of Parkinson’s disease using an adaptive fuzzy k-nearest neighbor approach”, Biomedical Signal Processing and Control, Elsevier, **(2013)**, pp. 364-373.

Rusdah and Edi Winarko, “Review on Data Mining Methods for Tuberculosis Diagnosis” Information Systems International Conference (ISICO), 2 – 4 December **2013.**

Marina Evrim Johnson and Nagen Nagarur, “Multi-stage methodology to detect health insurance claim fraud”, Health Care Management Science, DOI 10.1007/s10729-015-9317-3, Springer, 20 January **2015.**

Johannes Arthuur Govaert, Anne Charlotte Madeline van Bommel, Wouter Antonie van Dijk, Nicoline Johanneke van Leersum, Robertus Alexandre Eduard Mattheus Tollenaar and Michael Wilhemus Jacobus Maria Wouters, “Reducing Healthcare Costs Facilitated by Surgical Auditing: A Systematic Review”, Worls J Surg, DOI 10.1007/s00268-015-3005-9, Springer, 18 February **2015.**

Peng Z *et al.,* “An expression signature at diagnosis to estimate prostate cancer patients’ overall survival”, Prostate Cancer and Prostatic Disease (2014) 17, 81-90, doi 10.1038/pcan.2013.57; January **2014**, Nature.