IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

Medical Diagnosis Systems - Optimization Techniques

Mrs. R. Karthiyayini[1], Dr. R. Balasubramanian[2]

Assistant Professor, Department of Computer Application, Anna University (BIT Campus), Trichy[1]

Professor, P.G. and Research Department of Computer Science, JJ College of Arts and Science, Pudukkottai [2]

Abstract: Data mining is the general term for processing that is found in data schemas, usually with the help of powerful algorithms to automatically search for parts of the process. These methods come from disciplines such as statistics, databases, learning (IA) machines, pattern recognition, neural networks, visualization, high performance and parallel computing. The goal of data mining is that it is the conversion of data of facts or words that can be handled by computer knowledge. Today, dependency on data in medical care is increasing. The medical industry generates a large amount of data often in a large number of medical imaging. The availability of patient records, patient monitors, and a large number of medical data results in the need for data analysis to effectively extract useful knowledge from unstructured formats. Medical diagnosis is very important, but can also be accurate and effective to carry out complex tasks. The purpose of this paper is to provide the reader with an understanding of data mining and its importance in the medical system. In this paper, we will introduce the concept of data mining and its significance in medical systems, such as classification, clustering, association rules, regression and so on.

Keywords: Association Rules, Classification, Clustering, Data mining, Medical Diagnosis System, Regression.

I. INTRODUCTION

Rapid growth, massive data collection and storage in large and large data repositories has far exceeded our human understanding capabilities without powerful tools. A powerful, versatile tool is needed to automate the discovery of valuable information and data from vast amounts of data into systematic knowledge. This demand led to the birth of data mining.

Data mining is commonly used as a synonym for another term, KDD. The process of knowledge discovery is an iterative sequence of the following steps:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

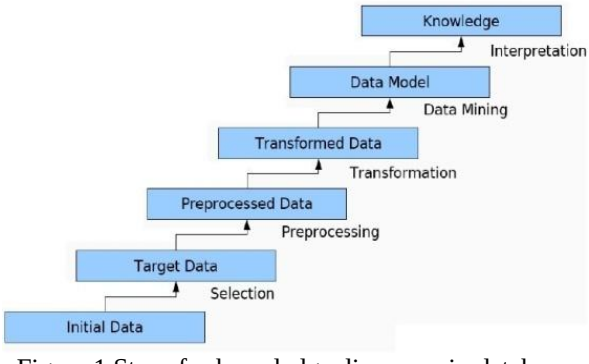

Figure 1 Steps for knowledge discovery in database

One of the most important steps in KDD is data mining. Data mining is the process of identifying new patterns and ideas of data. The insight gained through data mining can provide great value for making informed decisions. Data mining is classified in supervised learning (classification and prediction) and unsupervised learning (clustering and association rules, etc.).

There are a number of data mining functionalities. These include

1. Characterization and discrimination
2. The mining of frequent patterns, associations, and correlation.
3. Classification and regression
4. Clustering analysis and
5. Outlier analysis

The function of data mining is to specify the pattern found in the data mining task in the type. Under normal circumstances, these tasks can be divided into two categories.

1. Descriptive mining
2. Predictive mining

In descriptive mining can identify and summarize existing data and predict historical data mining is used to make predictions.

The process of data mining includes a hypothesis-making, data collection, pre-treatment, model estimation and model interpretation and conclusion.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

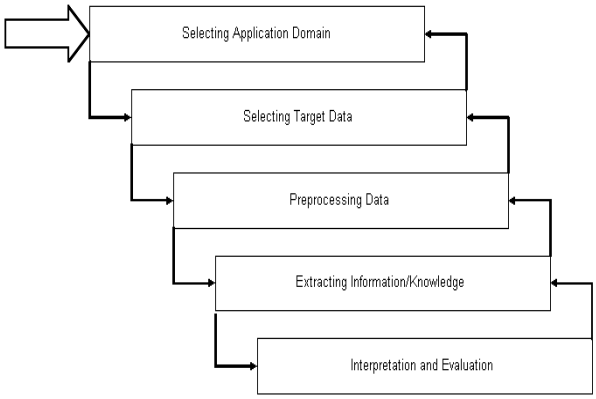

Figure 2 Steps for data mining Process

The content of this paper is organized as follows: In Section 2, some mining techniques are well known and widely used for data presentation. Section 3 briefly discusses the importance of data mining in medical systems. Section 4 summarizes the data mining techniques of medical systems.

II. DATA MINING TECHNIQUES

Some algorithms and data mining techniques include association, classification, clustering, regression, artificial neural network (ANN), decision tree, Bayesian classification, support vector machine (SVM) and many others. In this section, we propose that these four data mining techniques are widely used.

A. Association Rule Mining:
Association rules are the most dynamic and important knowledge model in data extraction [1]. The main focus of association rules is that associations in the quality domain can see certain needs. Pattern combination of events that occur concurrently. It provides a model for the organization of the patterns and methods of credit models often found. Frequent patterns and association rules are used to undermine knowledge of such scenarios.

In the diagnosis of association rules (AR) of the disease, the physician can support the treatment they can use for the patient. In fact the analysis of disease progression is not as easy as it may contain a fault diagnosis test and the presence of noise training samples. It uses an Apriori algorithm to recognize the conditions that often occur in a particular geographic area. A priori is a standard algorithm for extracting frequent itemsets. The process used to obtain the property Theapriori previous articles.

In the medical field, the use of the improved Apriori [2] discovery set is repeated in a directory of medical discovery algorithm terms, and makes the robust RA decree find out the association between the inference datasets or configure the huge ones. It shows that the improved method can extract the association rules of the articles and the drug list of the disease, which can help the expert in the medical analysis of the nature of the prototype. Data mining applications in grassland are DNA sequence analysis, clinical analysis of neural networks, medical association analysis, and disease diagnosis.

The system ant colony (ACS) is one of the most recent heuristic algorithms for combinatorial optimization [3], and this study uses a system ant colony to extract a large database to find association rules effectively. If the system can take into account a number of constraints, association rules are more efficiently generated. Therefore, this study presents a new method to extract the association rules database using ant colony algorithm. In addition, they have taken into account various constraints. Using the actual case, the National Health Insurance Research Database, the results show that the method can provide a more concise rule of priori methods. The calculation time is also reduced. Association rule mining is one of the most popular techniques in data mining, which aims to draw one of the related transactions between itemsets [4]. However, association rule mining often results in a very large number of discovery rules, leaving the database analyst with the task of running the association rules to find other interesting. This paper reviews the research area of association rule mining. Several techniques for discovering the efficiency of association rules have arisen. However, as the size of the database increases and the decision to produce efficient, targeted marketing, marketing basket, marketing catalog analysis, etc. in order to reduce the Apriori algorithm limits generate a large number of association rules in the method , The Apriori algorithm is first applied to generating frequent itemsets, and then frequent itemsets are used to generate association rules that is, rarely found in the database these patterns are usually considered boring and support measures removed. Such a pattern is called a rare type [5]. A rare pattern is that a set of elements or rules that are supported is less than the minimum supported threshold. However, more attention has been paid to the extraction of rare items, although it has an important use of negative association rules removed from the set of rare items (I) [6]

(Ii) Statistical disclosure of risk assessments in which anonymous census data are rare patterns that may lead to statistical information disclosure
Iii) Detection of fraudulent patterns in which rare patterns of financial or financial data may suggest fraud and IV behavior) Bioinformatics in data microagregaciones Rare patterns may suggest hereditary disease-related aberrant activities [7].

B. Classification technique:
Classification is a data mining technique (machine learning) that is used to predict data instances of group members. Several major types of classification methods include induction decision, k-nearest vizino Bayesian networks, case-based reasoning, genetic algorithms and fuzzy logic techniques.

SVM is an effective solution to classification problems, and no assumptions about data distribution and interdependence of the model-free approach. In

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

epidemiological studies and population health surveys, SVM technology has a much better function than traditional statistical methods, such as regression, especially when it comes to factors that have small effects (for example, data on multivariate risk profiles for potentially associative genomes And gene expression profiling) Limited understanding of the potential biological relationships between sample size and risk factors. This is particularly true in the case of common complex diseases where there are many risk factors, including the interaction of gene-gene and gene-environment interactions, which must be taken into account in predicting models to obtain sufficient discrimination. Our working principle provides a promising proof of SVM's ability to predict only a small fraction of the variables. This sapproach can be extended to include a large number of datasets, including many other variables, such as genetic biomarkers, such as derived data from different domains.

This paper deals with the Bayesian theorem and describes the role of the Bayesian network in medical dispute cases arising from the application of patients who have had a stroke as a result of invasive diagnostic tests. The allegation of negligence is based on the premise that it should use another (non-invasive) test because it has a lower risk. The case raises a number of widespread concerns and widespread application of the decision process in the medical industry, including informed consent, passive patient care when the ethical errors and the study of hot issues in the commitment to "true positive" positive.

An immediate concern is the best way to present the Bayesian parameter in a way that can be understood by people who are generally resistant to mathematical equations. The car is presented so that it does not require mathematical knowledge and understanding [8] to display a trivial representation of the Bayesian parameters visually. This approach supports a variety of alternatives, making all hypotheses both easy to understand and providing significant potential benefits in many areas of medical decision making.

Classification has been identified as a major problem in the emerging field of data mining. Over the years, there have been many research classification algorithms [10] analysis of classification technology performance evaluation [9] [11] [12] [13] [14], comparison and evaluation of different classification algorithms mineral data [15] [16] They solve real-world problems, such as in medicine, engineering, business and other fields of application

Two important indicators for data mining algorithm performance are the accuracy of classification / prediction and training time [17]. These metrics are useful for selecting the best algorithm for classification task / predictive data mining. Empirical studies on data mining for these performance indicators are scarce. Therefore, the purpose of this study is to determine how the data mining classification algorithm performs with the size of the input data. Three qualifying data mining algorithms, decision trees, neural network multilayer perceptions (MLP) and Bayesian Naïve - are subjected to different sized simulation data. It takes a training algorithm and its analysis of the accuracy of the ratings of different sizes of data. The results show that, compared to the MLP algorithm and the decision tree Naive Bayes, it takes less time to train the data, but the accuracy is lower.

K-Nearest Neighbor (KNN) is one of the most popular pattern recognition algorithms. Many researchers have found that the KNN algorithm performs very well in their experiments on different datasets. The traditional classification algorithm text KNN has three limitations: (i) computational complexity due to the use of all training samples for classification, (ii) the dependence on the training set and the performance (iii) there is now a difference between the samples. In order to overcome the limitations, KNN improved the version presented in. Genetic algorithm (GA) is combined with KNN to improve performance classification. Instead, consider all training samples and take K-Neighborhoods, using a genetic algorithm to immediately let K-Neighbor and then calculate the distance to the test samples to be classified.

C. Clustering technique

Clustering is the process of grouping similar data. Grouping techniques can be considered the most important learning under unsupervised, and any other type of problem of this type; try to find out the structure in the unlabeled data set. Partitioning a group of objects with uniform clusters [18,19] is the basic operation of data mining. The need for multiple data mining tasks such as unsupervised classification and data volume and a large number of heterogeneous datasets into smaller uniform subsets can be easily managed, manipulated separately and modeled for analysis. Clustering is a popular method for implementing this operation. The clustering method [20] groups a group of objects so that the objects in the same group are more similar to each other according to some criteria, different clusters of objects, and so on.

K-means clustering, clustering, clustering DBSCAN, optics, STING: This paper [21] analyzes six types of clustering techniques. The hierárquico cluster provides a clear understanding of the data classified visually by the dendrogram. The algorithm only links the agglomerate group, where the fuse is close to all the subgroups of the larger group, but leaves the disadvantage of smaller subgroups [22] from larger groups (called dispersion degrees). There is also a drawback that relatively large subgroups tend to absorb other subgroups during the melting process [23]. Clusters formed by this method (nearest neighbor) reach local proximity. All links of the same type use the farthest neighbors to merge adjacent clusters. Although the technique has overcome the degree of dispersion, it has been found that it is difficult to incorporate large populations, but to achieve global proximity [24]. UPGMA [25] exceeds the dispersion, and by using this set of averages the system approaches the equilibrium between the whole. A hierarchical clustering algorithm, such as a hybrid (BHHC), requires a priori

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

knowledge of the group's data. Parameters, such as the number of cluster centers and clusters, affect cluster partitioning [26]. The hierarchical clustering algorithm blockmodeling (HCUBE) uses a pair of objects to construct equivalence between them to identify similarity. Has been successfully applied in social networks to identify groups of different pages to achieve similarity in the cluster. If their network connection is the same [27], both pages of X and Y are considered structurally equivalent. These relationships are expressed in matrix relations and dissimilar matrices. The distance and interconnections between objects are determined using the Euclidean distance function or density in each group and are measured by the variance formula [28]. Sequence clustering algorithm based on hierarchical partitioning, the failure of the data sequence changes in different measures to maintain consistency, strictly partial order. The same method works well in the presence of split Missing and noisy data, but only provides binary division [23]. The hierarchical clustering algorithm based on sub-leader leadership is used to represent the leader (leader) and the group representative (sub-leader) [29]. This type of algorithm works in two stages and is computationally expensive as a means of finding the representation and then segmenting the data based on conventional clustering algorithms. Other algorithms such as PCTREE have high spatial complexity as the details of stored patterns and their characterization [30]. The most clustering algorithm correlates the correlation between objects, and then uses the classical clustering algorithm for partitioning and thus increasing the complexity of time [31]. The hierarchical relationships among the relevant clusters can be obtained by decomposing the data based on the spatial correlation in the higher dimension. The PCA is used to select clusters and the centroids of other objects belonging to group [32]. ACCA (Clustering Algorithm Mean Correlation), by averaging correlation measurements based on the data (genotype) similarity, the work is placed in k clusters [33]. Since the correlation between each pair of genes is calculated, the computation time is high for the increase of the number of genes. Based on the high correlation between objects in the same cluster, the correlation between any two data objects is used to find the correlations of similar clustering algorithms in their clusters and partitions. As the number of objects increases the complexity also increases, so there is a limitation in size.

Statistical clustering methods using the similarity measures, and according to the concept of conceptual clustering of objects carrying objects grouped object partition. Data mining applications usually classify data as contrary [34]. The greatest advantage of clustering can be said to resemble the identity of the type of object. By using clustering techniques, we can determine more dense and sparse areas in the object space, and we can find the general distribution patterns and relationships in the data attributes. Categorization can also be used as an effective means to distinguish between object groups or classes, but is expensive, so that grouping can be used as a preconditioning method for the selection and classification of attributes. For example, in order to form a customer group according to the genotype of the purchase pattern and the like.

Clustering method type
• Partitioning methods
• Method Hierarchical clustering (splitting)
• Density-based methods
• Web-based approach
• Model-based approach

Clustering is an unsupervised classification of patterns (observations, data elements, or eigenvectors) in groups (clusters). The caking problem is solved in many cases and in many disciplines by researchers. This reflects its broad appeal and practicality as a step in exploratory data analysis.
However, grouping is a difficult problem in combination and the assumptions and background differences in different communities make concepts and useful general methods very slow to convey. Our current clustering techniques categorize and identify cross-problems and recent developments. We also describe some important applications of clustering algorithms, such as image segmentation, object recognition and retrieval.

In particular, this paper reveals that the problem of cervical cancer diagnosis is the way in which the mining analyst learns in the machine context [35]. So far, it is not used in data mining techniques such as clustering to analyze cervical cancer patients. Therefore, attempts were made to identify patterns in the database of cervical patients using clusters.
In the prediction-means of cardiac disease through clustering techniques KA [36], the following step1 continues. Ķ indicates that the data points to be grouped are placed in space. These points represent the chiefs of the main group. 2. The data is allocated to groups of adjacent hundred. 3. All K centroid locations, once all data are allocated for recalculation. 4. Steps 2 and 3 are repeated until the center of gravity stops moving. This leads to a deliberate segregation of the data set that can minimize the measurement. Pretreatment of heart disease data using K-means clustering with K-means clustering is a multivariate statistical analysis also known as cluster analysis, categorical analysis unsupervised or numerical classification. The K-means clustering generates disjoint clusters, the number of planes (non-hierarchical). It is very suitable for producing globular clusters. The K-means method is digital, unsupervised, uncertain and iterative.
The diffuse contrast allows packets to belong to data points in more than one group. The resulting partitions, therefore, are diffuse partitions. Each group is associated with a membership function representing the extent to which each data point belongs to the group. In all methods of fuzzy clustering, Fuzzy C-Means (FCMC) is still dominant because of the successful application in academia and industry. Fuzzy C-Means Clustering An iterative search for a set of fuzzy clustering and clustering

DOI 10.17148/IARJSET.2016.31109

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

centers represents the optimal data structure for the partner. The algorithm is based on the number of clusters that the user specifies to be grouped in the data set. In general, the use of different distance functions or a slight modification of the objective function leads to clustering to be able to detect different types of clustering algorithms. Using the Euclidean distance, FCMC is able to detect spherically shaped clusters of similar size.

D. Regression technique**:**

Regression techniques can be applied to preaching. Regression analysis can be modeled using the relationship between one or more independent and dependent variables. In data mining, the independent variables are known to the attribute and the response variable is what we expected. Unfortunately, many real-world problems are not just predictions. For example, the prices of sales, shares, and product failure rates are difficult to predict because they rely on the complex interactions of multiple predictors. Therefore, more complex techniques (eg, Logistic Regression, Decision Tree, or Neural Network) may need to predict future values. The same type of model can often be used for two regressions as a classification. For example, the decision tree algorithm CART (classification tree and regression tree) can be used to construct a classification tree (response to categorical variable categorization) as a regression tree (continuous variable to predict the response). Neural networks can also create regression models for classification. The type of regression method.

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

[37] The purpose of statistical evaluation of medical data is to often describe the relationship between two variables or between variables. For example, one would like to know not only if the patient has high blood pressure, but if the possibility of having hypertension is influenced by factors such as age and weight. The dependent variable (blood pressure) is called the dependent variable, or the response variable; the explanatory variable (age, weight) is called the independent or predictor variable. The association's measures provide a preliminary impression of the degree of statistical dependence between variables. If the variable and the independent variable are continuous, such as blood pressure and body weight, then a correlation coefficient can be calculated as a measure of the intensity of the relationship between them. Regression analysis is a statistical evaluation that allows three things:

Description: The relationship between the dependent variable and the dependent variable can be from statistical regression analysis.
Estimate: The value of the dependent variable can be estimated from the observations Independent variable.
Prognostication: The risk factors that influence the outcome can be identified and the prognosis of the individual can be determined.

Regression analysis uses a model that describes the relationship between the dependent variable and the independent variable in a simplified mathematical form. It is probable that the biological reason is expected to be a priori for that particular type of mathematical function that best describes the relationship, or a simple hypothesis that must be raised that is the case (for example, blood pressure increases linearly with age).

[38] The analysis of receiver operating characteristics (ROC) is a useful method for assessing the accuracy of diagnosis. Development of ROC Hierarchical Bayesian Model Based on Nonlinear Regression Analysis. Conducted an analysis of prospective data diagnostic accuracy validation using a multicenter clinical trial of prostate cancer biopsy collected at three centers. The gold standard is based on radical prostatectomy to identify local and late disease. To assess the diagnostic performance of PSA levels at a fixed level of Gleason score, normal transformations were applied to the outcome data. The area of the area under the ROC curve (AUC) was estimated by stratification regression analysis for the effects of the cluster (clinical center) and the risk of cancer (low, intermediate and high).

The objective of this work [39] was to evaluate the implementation of the Logistic Regression Multinómica to assess the factors affecting underweight and overweight of women. This is a cross-sectional study. With a detailed record of the socio demographic of women with height and weight. Simple random sampling in the selection of females was obtained. Pregnant women were excluded from the study. Women's weight status can be indicated by BMI category indicators used in the analysis of outcome variables. A total of 435 women were interviewed. In order to assess the net effect of exposure measures on outcome measures, polynomial Lo-gistic regression analysis is considered appropriate because the results are inherently multifarious. The conclusion was that the majority of women were between the ages of 20 and 30 (44.4%). More than a third of women have rupee income between families. 5001-10, 000 per month (57.7 per cent) and illiteracy (68.5 per cent). Polynomial regression results show that overweight increases with age and education is higher among urban women with higher morbidity.

III.APPLICATION OF DATA MINING IN MEDICAL SYSTEM

Application areas for data mining techniques in medical systems include disease diagnosis and prognosis, discovering frequent patterns mining in specific diseases, analysis of patient's medical records, etc. In the health field, data mining applications have been growing considerably as it can be used to directly derive patterns, which are relevant to forecast different risk groups among the patients.

Data Mining acting an important role in the Prediction of Cancer Diseases. The data mining methods judgment were aim as a main objective in many studies that mainly

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

targeted to develop a prediction model in a dangerous fields, like medicine, by examining several data mining methods, proposed to get the model that have the highest prediction accuracy.

More deadly than breast, cervical, and prostate cancer, it has been estimated that oral cancer kills one person every hour, every day **[40]**. In this studies proposed that head and neck cancer and tongue cancer in specific is growing in early adults. Oral cancer is a usually recognized type of head and neck cancer, which is increasing globally in occurrence and growing critically in many regions of the countries in the world. Most important step in reducing the death rate from oral cancer is early diagnosis.

Coronary heart disease (CHD) is a serious disease that causes many deaths especially in China. The study on CHD patients was aimed to identify the syndrome of CHD using data mining techniques [41]. The study used 1069 CHD cases that were collected using surveys on 5 clinical centres located in two provinces. 80 symptoms that are closely related to CHD and always appear in the literatures of CHD were selected.

The study of breast cancer is interesting because the risk factors are difficult to identify similar to childhood obesity. The study on breast cancer recurrences involve 1035 breast cancer patient **[42]**. 22 medical patient features were recorded at the time of surgery while 10 more features were recorded through follow-up. The study took more than 10 years.

Diabetes is a metabolic disorder where the body cannot make proper use of carbohydrate and greatly affected by the patient lifestyle **[43, 44]**. The study on diabetes prediction used 2017 diabetic patient clinical information **[43]**. There are 425 features in the database. The first step in the study was data pre-processing: data integration and reduction. The following step was feature selection using Relief to reduce the number of parameters.

The prediction was to identify the number of embryos to be transferred to the woman's womb and the selection of embryos with the highest reproductive viabilities **[45]**. The prediction was to determine whether the embryos are suitable for implant. The similarity of IVF implantation prediction with childhood obesity prediction is the imbalances distribution of positive and negative samples, which is common in medical datasets **[45].**

Six data mining techniques and logistic regression were used for childhood obesity predictions **[46]**. The techniques are decision tree (C4.5), association rules, Neural Network, Naïve Bayes, Bayesian networks, linear SVM and RBF SVM. The prediction aims to identify obese and overweight children at 3 years old using the data recorded at birth, 6 weeks, 8 months, and 2 years. The prediction used 16653 instances, where only 20% of the samples are obese or overweight cases. The accuracy was measured using the sensitivity and specificity.

IV. CONCLUSION

This paper presents a review of data mining importance in medical systems. Data mining is the process of analyzing and summarizing data from different perspectives and converting it into useful information.
About ¾ billion of people's medical records are electronically available. Data mining in medicine distinct from other fields due to nature of data such as heterogeneous, with ethical, legal and social constraints.

Well-known data mining techniques include the Artificial Neural Network (ANN), decision tree, Bayesian classifiers, Support Vector Machine (SVM). Data mining utilization is increasing in medical informatics and for improving the decision making such as diagnostic and prognostic problems in oncology, liver pathology, Neuropsychology, and Gynecology. Medical data mining can be the most rewarding despite the difficulty.

REFERENCES

[1] J. Y Wang, H. Y Wang and D. W Zhang, et al, "Research on Frequent Itemsets Mining Algorithm based on Relational Database", Journal of Software, vol. 8, no. 8, pp. 1843-1850, 2013.
[2] Kamran Shaukat, Sana Zaheer, Iqra Nawaz,"Association Rule Mining: An Application Perspective", International Journal of Computer Science and Innovation Vol. 2015, no. 1, pp. 29-38 ISSN: 2458-6528
[3] R.J. Kuo_, C.W. Shih, "Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan"
[4] Rajdeep Kaur Aulakh, "Association Rules Mining Using Effective Algorithm: A Review",International Journal of Advanced Research in Computer Science and Software Engineering,Volume 5, Issue 3, March 2015 ISSN: 2277 128X
[5] D. J. Haglin and A. M. Manning,"On minimal Infrequent itemset mining", In DMIN, pages 141–147, 2007.
[6] X. Wu, C. Zhang, and S. Zhang, "Efficient mining of both positive and negative association rules", ACM Transactions on Information Systems, 22(3):381–405, 2004.
[7] X. Dong, Z. Niu, X. Shi, X. Zhang, and D. Zhu, " Mining both positive and negative association rules from frequent and infrequent itemsets" In ADMA, pages 122–133,2007.
[8] "Comparing risks of alternative medical diagnosis using Bayesian arguments", Journal of Biomedical Informatics.
[9] Raj, K. and Rajesh, V. (2012) "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET).
[10] Pardeep, K., Nitin, V.K. and Sehgal, D.S.C. (2012),"A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems", International Journal of Data Mining & Knowledge Management Process (IJDKP).
[11] Jyoti, S., Ujma, A., Dipesh, S. and Sunita, S. (2011)," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction". International Journal of Computer Applications.
[12] S. OlalekanAkinola, O. JephtharOyabugbe, "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study" Journal of Software Engineering and Applications, 2015, 8, 470-477 Published Online September 2015 in SciRes.
[13] M. and Heckerman, D. (February, 1998), " An experimental comparison of several clustering and initialization methods",Technical Report MSRTR-98-06, Microsoft Research, Redmond,WA.
[14] Sharma,N. , Bajpai,A., and Litoriya, R. 2012,,"Comparison the various clustering algorithms of weka Tools", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 5, May 2012.

[15] N. Ailon, M. Charikar, and A. Newman. (2005)"Aggregating inconsistent information: ranking and clustering". In Proceedings of the thirtyseventh annual ACM Symposium on Theory of Computing, pages 684–693, 2005.
[16] Thirunavukkarasu, K.S. and Sugumaran, S. (2013), "Analysis of Classification Techniques in Data Mining." IJESRT: International Journal of Engineering Sciences & Research Technology, 3640-3646.
[17] Abirami, N., Kamalakannan, T. and Muthukumaravel, A. (2013)," A Study on Analysis of Various Data Mining Classification Techniques on Healthcare Data", International Journal of Emerging Technology and Advanced Engineering.
[18] Liu, Y., Pisharath, J., Liao, W.-K., Memik, G., Choudhary, A. and Dubey, P. (2002), "Performance Evaluation and Characterization of Scalable Data Mining Algorithms". Intel Corporation, CNS-0406341.
[19] Nikhil, N.S. and Kulkarni, R.B. (2013),"Evaluating Performance of Data Mining Classification Algorithm in Weka". International Journal of Application or Innovation in Engineering & Management.
[20] Daniela, X., Christopher, J.H. and Roger, G.S. (2009), "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages". IJSCI: International Journal of Computer Science Issues.
[21] "A Review: Comparative Study of Various Clustering Techniques in Data Mining" v313-0162
[22] Yu L, Gao L, Li K, Zhao Y and Chiu D.K.Y," A degree-distribution based hierarchical agglomerative clustering algorithm for protein complexes identification, Computational Biology and Chemistry", vol. 35, 2011, pp 298 - 307.
[23] Lee John W.T, Yeung D.S, Tasng,"E.C.C: Hierarchical clustering based on ordinal consistency, Pattern Recognition", vol. 38, 2005, pp 1913-1925.
[24] Han J., Kamber, M. and Pei J: "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.
[25] Wu J, Xiong H and Chen J," Towards understanding hierarchical clustering: A data distribution perspective, Neuro Computing", vol. 72, 2009, pp 2319 - 2330.
[26] Vijaya P.A, NarasimhaMurty and Subramanian," Leaders-Subleaders: An efficient hierarchical clustering algorithm for large data sets, Pattern Recognition Letters", vol. 25, 2004, pp 505-513.
[27] Qiao S, Li Q, Li H, Peng J and Chen H, "A new block modeling based hierarchical clustering algorithm for web social networks, Engineering Applications of Artificial Intelligence", vol. 25, 2012, pp 640 - 647.
[28] Tu Q, Lu J.F, Yuan B, Tang J.B and Yang J.Y," Density based Hierarchical Clustering for streaming data, Pattern Recognition Letters", vol. 33, 2012, pp 641 - 645.
[29] Vijaya P.A, NarasimhaMurty and Subramanian D.K," Efficient bottom-up hybrid hierarchical clustering techniques for protein sequence classification, Pattern Recognition", vol. 39, 2006, pp 2344 - 2355.
[30] Aqueel Ahmed, Shaikh Abdul Hannan,"Data Mining Techniques to Find Out Heart Diseases: An Overview",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012
[31] Astrid Schneider, Gerhard Hommel, and Maria Blettne "Linear Regression Analysis", Part 14 of a Series on Evaluation of Scientific Publications
[32] Kelly H. Zou1,2,* and A. James O'Malley1 "A Bayesian Hierarchical Non-Linear Regression Model in Receiver Operating Characteristic Analysis of Clustered Continuous Diagnostic Data " NIH Public Access
[33] Shivam Dixit1, Bhushan Kumar2, Abhay Singh3, Ramkumar Ashoka4 "An Application of multinomial Logistic Regression to Assess the Factors Affecting the Women to Be Underweight and Overweight: A Practical Approach **"**International Journal of Health Sciences & Research Vol.5; Issue: 10; October 2015
[34] "Unconscious Oral Cancer Detection using Data Mining Classification Approaches"
[35] Ananthanarayana V.S., NarasimhaMurty and M., Subramanian, D.K.: "Efficient clustering of large data sets, Pattern Recognition", vol.34, 2001, pp 2561-2563.
[36] Seal S, Komarina S and Aluru S, An optimal hierarchical clustering algorithm for gene expression data, Information Processing Letters, vol. 93, 2005, pp 143-147.
[37] Zimek A, Thesis on Correlation clustering, University of Munchen, 2008.
[38] Bhattacharya A and De, Rajat K.: Average correlation clustering algorithm (ACCA) for grouping co-regulated genes with similar pattern of variation in their expression values, Journal of Biomedical Informatics, vol.43, 2010, pp560-568.
[39] Pankaj Saxena&SushmaLehri,"ANALYSIS OF VARIOUS CLUSTERING ALGORITHMS OF DATA MINING ON HEALTH INFORMATICS"
[40] "Data mining approach to cervical cancer patients analysis using clustering technique", Asian journal of information technology medwell online 2006