Optimization techniques for Medical Diagnosis Systems

Dr. R. Balasubramanian[1], Mrs. R. Karthiyayini[2]
[1, 2] Department of Computer Application
[1] Karpaga Vinayaga College of Engineering and Technology, chengalpet
[2] Anna University(BIT Campus) , Trichy

Abstract- *Data mining is an umbrella term referring to the process of discovering patterns in data, typically with the aid of powerful algorithms to automate part of the search. These methods come from the disciplines such as statistics, data bases, machine learning (AI), pattern recognition, neural networks, visualization, high-performance and parallel computing. The goal of data mining is to turn data that are fact, numbers, or text which can be processed by a computer into knowledge. Nowadays, the reliance of health care on data is increasing. The medical industry generates huge amounts of data often exists in vast quantities in an unstructured format from medical records, patient monitoring, and medical imaging. The availability of huge amounts of medical data leads to the need for powerful data analysis to extract useful knowledge. Medical diagnosis is extremely important but complicated task that should be performed accurately and efficiently. Various data mining techniques exist for disease diagnosis in healthcare industry namely classification, clustering, association rules, regression etc., Therefore, this paper aims to allow the readers to understand about data mining and its importance in medical systems.*

I. INTRODUCTION

The fast growing, tremendous amounts of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. Data mining as a synonym for another popularly used term, KDD. The knowledge discovery process is 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

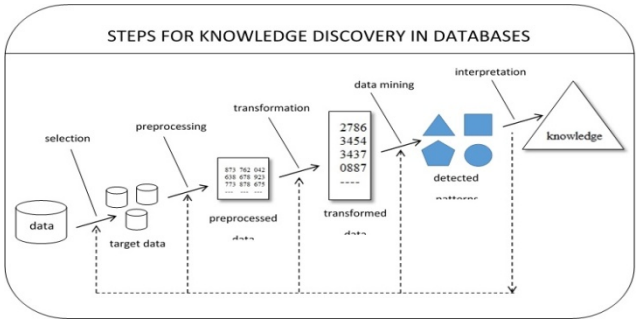

Figure 1

One of the most important steps of the KDD is the data mining. Data mining is the process of identifying new patterns and insights in data. Insight derived by data mining can provide tremendous value, for making informed decision. Data mining can be classified into 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

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories.

1. Descriptive mining
2. Predictive mining

In descriptive mining can identify and summarize the already existing data and in predictive mining historical data is used to make prediction.

The data mining processes include formulating a hypothesis, collecting data, performing preprocessing, estimating the model, and interpreting the model and draw the conclusions.

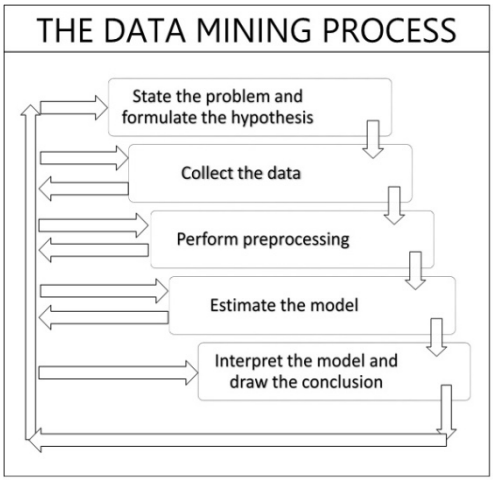

Figure 2

This paper is organized as follows. In Section 2, several well-known and widely used data mining techniques are presented. Section 3 briefly discusses the importance of data mining in medical systems. Section 4 presents the summary on the study of the data mining techniques for medical systems.

II. DATA MINING TECHNIQUES

Various algorithms and techniques in data mining include the Association, Classification, Clustering, Regression, Artificial Neural Network(ANN), decision tree, Bayesian Classifiers, Support Vector Machine(SVM) and many others. In this section we introduce these four widely used data mining techniques.

Association Rule Mining:

Association rule is the most dynamic and important knowledge models of data extraction **[1]**.The main focusof association rule is on the association of the quality domain that can see the certain necessities. The patterns exposethe combination of the events that occur at the same time. It provides well-organized method of frequent pattern discovery and model credit. Frequent patterns and association rules are the knowledge that used to mine in such a scenario.

In disease diagnosis association rules(AR) can be used that will be helpful for supporting physicians to treat patients. Practically disease analysis is not an easy progression as it may contain faulty diagnosis test and the presence of the noise in training examples. Apriori algorithm has been used to recognize frequently occurring diseases in particular geographic area. Apriori is the most standard algorithm for extracting frequent item sets. The a priori procedure uses previous facts of itemset belongings.

In medical field, using the improved Apriori algorithm **[2]**to find the recurrent element sets in a catalogue of the medicinal finding, and makes the robust AR in edict to find out inference association or configurations among the huge data item sets. It shows that the modified apriori procedure can take out the association rule prototypes about belongings and nature of the sickness from the medicinal catalogues, which can help specialists in medical analysis. In the meadow of medicinal the data excavating applications are DNA sequence analysis, neural network in clinical analysis, association analysis in medicine, disease diagnosis.

The ant colony system (ACS) is one of the newest meta-heuristics for combinatorial optimization problems**[3]**, and this study uses the ant colony system to mine a large database to find the association rules effectively. If this system can consider multi-dimensional constraints, the association rules will be generated more effectively. Therefore, this study proposes a novel approach of applying the ant colony system for extracting the association rules from the database. In addition, the multi-dimensional constraints are taken in to account. The results using a real case, the National Health Insurance Research Database, show that the proposed method is able to provide more condensed rules than the Apriori method. The computational time is also reduced.

Association rule mining is one of the most popular techniques of data mining methods whose aim is to extract associations among sets of items in transaction databases **[4]**. However, mining association rules often results in a very large number of found rules, leaving the database analyst with the task to go through all the association rules and discover interesting ones. Authors reviewed mining association rules has attracted a lot of attention in the research community. Several techniques for efficient discovery of association rules have appeared. However, with the increase in the size of the databases and for efficient decision making, selective marketing, market basket analysis, catalogue marketing industry etc. To reduce the limitation of Apriori algorithm of generating large number of association rules, an algorithm was proposed. In the proposed method, initially applied Apriori algorithm in order to generate frequent item-sets and then frequent item-sets are used to generate association rules.

Patterns that are rarely found in database are often considered to be uninteresting and are eliminated using the support measure. Such patterns are known as infrequent patterns [5].An infrequent pattern is an item set or a rule whose support is less than the minimum support threshold. However, significantly less attention has been paid to mining of infrequent item sets, even though it has got important usage

in (i) mining of negative association rules from infrequent item sets [6],
(ii) statistical disclosure risk assessment where rare patterns in anonymous census data can lead to statistical disclosure
(iii) fraud detection where rare patterns in financial or tax data may suggest unusual activity associated with fraudulent behavior and (iv) bioinformatics where rare patterns in micro array data may suggest genetic disorders[7].

Classification technique:

Classification is a data mining (machine learning)technique used to predict group membership for data instances. Several major kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques.

In this paper**[8]** explains the role of Bayes Theorem and Bayesian networks arising in a medical negligence case brought by a patient who suffered a stroke as a result of an invasive diagnostic test. The claim of negligence was based on the premise that an alternative (non-invasive) test should have been used because it carried alower risk. The case raises a number of general and widely applicable concerns about the decision-making process within the medical profession, including the ethics of informed consent, patient care liabilities when errors are made, and the research problem of focusing on 'true positives' while ignoring 'false positives'.
An immediate concern is how best to present Bayesian arguments in such a way that they can be understood by people who would normally balk at mathematical equations. Authors proposed to present purely visual representations of a non-trivial Bayesian argument in such a way that no mathematical knowledge or understanding is needed**[8]**. The approach supports a wide range of alternative scenarios, makes all assumptions easily understandable and offers significant potential benefits to many areas of medical decision-making.

Classification has been identified as an important problem in the emerging field of data mining. Over the years, there has been quite a number of tremendous studies on classification algorithms **[9] [10]**, analysis of classification techniques**[11] [12]**, performance evaluation**[13] [14]**, comparisons and evaluations of different data mining classification algorithms **[15] [16]** alongside their applications in solving real world problems such as in theare as of medicine, engineering, business etc.

Two important performance indicators for data mining algorithms are accuracy of classification/prediction and time taken for training**[17]**. These indicators are useful for selecting best algorithms for classification/prediction tasks in data mining. Empirical studies on these performance indicatorsin data mining are few. Therefore, this study was designed to determine how data mining classification algorithm perform with increase in input data sizes. Three data mining classification algorithms—Decision Tree, Multi-Layer Perception (MLP) Neural Network and Naïve Bayes—were subjected to varying simulated data sizes. The time taken by the algorithms for trainings and accuracies of their classifications were analyzed for the different data sizes. Results show that Naïve Bayes takes least time to train data but with least accuracy as compared to MLP and Decision Tree algorithms.

k-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition. Many researchers have found that the KNN algorithm accomplishes very good performance in their experiments on different data sets. The traditional KNN text classification algorithm has three limitations: (i) calculation complexity due to the usage of all the training samples for classification, (ii) the performance issolely dependent on the training set, and (iii) there is no weight difference between samples. To overcome these limitations, an improved version of KNN was proposed. Genetic Algorithm (GA) was combined with KNN to improve its classification performance. Instead of considering all the training samples and taking k-neighbors, the GA is employed to take k-neighbors straightaway and then calculate the distance to classify the test samples.

Clustering technique:

Clustering is a process of putting similar data into groups. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data.

Partitioning a set of objects into homogeneous clusters **[18,19]** is a fundamental operation in data mining. The operation is needed in a number of data mining tasks, such as unsupervised classification and data summation, as well as segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed. Clustering is a popular approach used to implement this operation. Clustering methods **[20]** partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria.

In this paper**[21]** reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBSCAN clustering, OPTICS , STING. Hierarchical clustering provides a clear understanding of the classified data pictorially through a dendogram. Single link algorithm is an agglomerative clustering that merges all the sub clusters closer to the larger cluster but has the drawback of leaving smaller sub clusters away from the larger cluster (called dispersion degree) **[22]**. It also has a disadvantage that the relatively larger sub clusters tend to absorb the other sub-clusters during the merging process **[23]**. The clusters formed by this method (nearest neighborhood) achieve local proximity. The complete linking of the same type uses farthest neighborhood to merge the neighbor clusters. Though this technique overcomes dispersion degree, it finds difficulty in merging larger clusters but achieves global proximity **[24]**. UPGMA **[25]** overcomes the tradeoff between dispersion degree and global proximity by using group average scheme. Algorithms like Hybrid hierarchical clustering (BHHC) requires prior knowledge about the data for clustering. The parameters like cluster center and number of clusters influence the cluster partitioning **[26]**. The Block modeling hierarchical clustering algorithm (HCUBE) uses structural equivalence between a pair of objects to identify the similarity between them. It has been successfully applied in social networking to identify the distinct cluster of pages to achieve the similar pages in a cluster. Two pages X and Y are considered to be structurally equivalent if their connections to the network are identical **[27]**. These relations are represented in relation matrix and Dissimilarity matrix. The closeness and inter-connectivity between the objects are identified using Euclidean distance or density function with in each cluster and is measured through variance formula **[28]**. Ordinal consistency clustering algorithm preserves the strict partial ordering based on the dissimilarity measures for hierarchical partitioning, which fails when the data ordering changes. The divisive order invariant approach works effectively in presence of missing and noisy data but only provides binary split **[23]**. The Leaders – Sub leaders based hierarchical clustering algorithm uses representatives (leaders) and sub representatives (sub leaders) of the clustering **[29]**. This sort of algorithms work in two stages like finding representatives and then partitioning the data based on conventional clustering algorithm and is computationally expensive. Other algorithms like PC Tree have high space complexity as they store the details of patterns and their representations **[30]**. Most of correlation clustering algorithms compute the correlation between the objects and then use conventional clustering algorithms for partitioning and hence increase the time complexity **[31]**. Hierarchical Relationship between correlation clusters can be obtained by decomposing the data based on correlation in high dimensional space. Eigen values, PCA are used to select the centroid of the cluster and other objects belonging to the cluster **[32]**. ACCA (Average correlation clustering algorithm) works based on the basis of similarity of the data (gene) measured through average correlation are put into k clusters **[33]**. Since correlation between each pair of genes is computed the computational time is high as number of genes grows. The correlation clustering algorithms use correlation between any two data objects to find their similarity in the cluster and the partitions are formed based on high correlation between objects in the same cluster. As number of objects increases the complexity also increases, thus there are limitations on dimensions.

Statistical clustering methods use similarity measures to partition objects whereas conceptual clustering methods cluster objects according to the concepts objects carry. Data mining applications frequently involve categorical data**[34]**. The biggest advantage of Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative(divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines. This reflects its broad appeal and usefulness as one of the steps in exploratory data analysis.

However, clustering is a difficult problem combinatorically, and differences in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur. We present taxonomy of clustering techniques, and identify cross-cutting themes and recent advances. We also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

In particular, the article reveals that how the problem of cervical cancer diagnosis is approached by a data mining analyst with a background in machine learning**[35]**. To this point data mining technique such as clustering has not been used to analyze cervical cancer patients. Hence, made an attempt to identify patterns from the database of the cervical patients using clustering.

In a prediction of heart disease using K – Means clustering technique**[36],**the following steps are followed1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centurions. 2. The data are assigned to the group that is adjacent to the centurion. 3. The positions of all the K centroids are recalculated as soon as all the data are assigned. 4. Steps 2 and 3 are reiterated until the centroids stop moving any further. This results in the segregation of data into groups from which the metric to be minimized can be deliberated. The preprocessed heart disease data is clustered using the K-means algorithm with the K values. Clustering is a type of multivariate statistical analysis also known as cluster analysis, unsupervised classification analysis, or numerical taxonomy. K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative.

Regression technique:

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models. Types of regression methods

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

[37]The purpose of statistical evaluation of medical data is often to describe relationships between two variables or among several variables. For example, one would like to know not just whether patients have high blood pressure, but also whether the likelihood of having high blood pressure is influenced by factors such as age and weight. The variable to be explained (blood pressure) is called the dependent variable, or, alternatively, the response variable; the variables that explain it (age, weight) are called independent variables or predictor variables. Measures of association provide an initial impression of the extent of statistical dependence between variables. If the dependent and independent variables are continuous, as is the case for blood pressure and weight, then a correlation coefficient can be calculated as a measure of the strength of the relationship between them. Regression analysis is a type of statistical evaluation that enables three things:

Description: Relationships among the dependent variables and the independent variables can be statistically described by means of regression analysis.

Estimation: The values of the dependent variables can be estimated from the observed values of the independent variables.

Prognostication: Risk factors that influence the outcome can be identified, and individual prognoses can be determined.

Regression analysis employs a model that describes the relationships between the dependent variables and the independent variables in a simplified mathematical form. There may be biological reasons to expect a priori that a certain type of mathematical function will best describe such a relationship, or simple assumptions have to be made that this is the case (e.g., that blood pressure rises linearly with age).

[38]Receiver operating characteristic (ROC) analysis is a useful evaluative method of diagnostic accuracy. A Bayesian hierarchical nonlinear regression model for ROC analysis was developed. A validation analysis of diagnostic accuracy was conducted using prospective multi-center clinical trial prostate cancer biopsy data collected from three participating centers. The gold standard was based on radical prostatectomy to determine local and advanced disease. To evaluate the diagnostic performance of PSA level at fixed levels of Gleason score, a normality transformation was applied to the outcome data. A hierarchical regression analysis incorporating the effects of cluster (clinical center) and cancer

risk (low, intermediate, and high) was performed, and the area under the ROC curve (AUC) was estimated.

The objective of the paper[39] is to evaluate the application of Multinomial Logistic Regression to assess the factors affecting the women to be underweight and overweight. This was a cross-sectional study. The socio-demographic detail with height and weight of the women was recorded. The simple random sampling was adopted in the selection of women. The pregnant women were excluded from the study. Women's weight status, indicated by their BMI category, was used as the outcome variable in the analyses. A total of 435 women were interviewed. To assess the net effect of exposure variables on the outcome measures, multinomial logistic regression analysis was contemplated to be suitable as the outcome measure is polychotomous by nature. It concludes most of the women were between 20-30 years of age (44.4%). More than one third of the women had family income between Rs. 5001-10, 000 (57.7%) per month and were illiterates (68.5%). The results of multinomial logistic regression showed that overweight increased with age and education with higher prevalence among urban women.

III. APPLICATION OF DATA MINING IN MEDICAL SYSTEMS

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 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] Thirunavukkarasu, K.S. and Sugumaran, S. (2013), "Analysis of Classification Techniques in Data Mining." IJESRT: International Journal of Engineering Sciences & Research Technology, 3640-3646.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] M. and Heckerman, D. (February, 1998), " An experimental comparison of several clustering and initialization methods",Technical Report MSRTR-98-06, Microsoft Research, Redmond,WA.

[19] 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.

[20] 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.

[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] Ananthanarayana V.S., NarasimhaMurty and M., Subramanian, D.K.: "Efficient clustering of large data sets, Pattern Recognition", vol.34, 2001, pp 2561-2563.

[31] 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.

[32] Zimek A, Thesis on Correlation clustering, University of Munchen, 2008.

[33] 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.

[34] Pankaj Saxena&SushmaLehri,"ANALYSIS OF VARIOUS CLUSTERING ALGORITHMS OF DATA MINING ON HEALTH INFORMATICS"

[35] "Data mining approach to cervical cancer patients analysis using clustering technique", Asian journal of information technology medwell online 2006

[36] 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

[37] Astrid Schneider, Gerhard Hommel, and Maria Blettne "Linear Regression Analysis", Part 14 of a Series on Evaluation of Scientific Publications

[38] 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

[39] 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

[40] "Unconscious Oral Cancer Detection using Data Mining Classification Approaches"

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

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

[43] Yue Huang, et al., "Evaluation of outcome prediction for a clinical diabetes database ", ed, 2004.

[44] Y. Huang, et al., "Evaluation of Outcome Prediction for a Clinical Diabetes Database, Knowledge Exploration in Life Science Informatics." vol. 3303, J. López, et al., Eds., ed: Springer Berlin / Heidelberg, 2004, pp. 181-190.

[45] A. Uyar, et al., "ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction, Electronic Healthcare." vol. 27, P. Kostkova, Ed., ed: Springer Berlin Heidelberg, 2010, pp. 108-11

[46] S. Zhang, et al., "Comparing data mining methods with logistic regression in childhood obesity prediction," Information Systems Frontiers, vol. 11, p. 51, 2009.