Taxonomy of Ageing and Non-Ageing Genes by Means of General Data Mining Techniques

B.Madasamy
Asst. Prof & Research Scholar, Agni College of Technology, Anna University, Chennai

J.Jebamalar Tamilselvi, PhD
Director & Prof, Jaya Engineering College, Anna University, Chennai

ABSTRACT

Classification of DNA repair genes into ageing and non-ageing is a vital process to identify faulty genes. Classifying genes into ageing and non-ageing human genome ranges over ten thousand. The ratio of ageing genes in the human genome is very less. There is a need for classifying ageing genes accurately in order to understand the complex processes occurring in living organisms. Data mining approach is routinely applied to classify DNA repair genes using various characteristics and feature. This paper proposes to build classification models that allow us to discriminate between ageing-related and non-ageing related DNA repair genes, in order to enhance value their different properties of genes classification performance should be evaluated by applying different kinds of classification algorithms like pruning, multiperceptron and Logistics. It will helpful for biomedical researchers, gene analyzer, patients and different kinds of end user.

Keywords

Ageing, Genome, Pruning, Gene.

1. INTRODUCTION

Human longevity is a complex phenotype that has a significant genetic inclination. Among the biological processes, ageing process is governed through the regulation of signaling pathways and transcription factors. The DNA damage concepts of ageing suggest that ageing is a consequence of unrepaired DNA damage accumulation. Intensive research has been carried out to elucidate the role of DNA repair systems in the ageing and non-ageing process. A gene is a molecular unit of heredity of a living organism. It is a name given to some stretches of DNA and RNA that code for a polypeptide or for an RNA chain that has a function in the organism. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes contain information to build and maintain an organism's cells and pass genetic traits to offspring. All organisms have many genes corresponding to various biological traits, some of which are immediately visible, such as eye color or number of limbs, and some of which are not, such as blood type, increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life. A gene is the basic instruction sequence of nucleic acids (DNA or, in the case of certain viruses RNA), while an allele is one variant of that gene. In most cases, all people would have a gene for the trait in question, but certain people will have a specific allele of that gene, which results in the trait variant. Further, genes code for proteins, which might result in identifiable traits, but it is the gene, not the trait, which is inherited. [1, 9, 24]

Deoxyribonucleic acid (DNA) molecules are informational molecules encoding the genetic instructions used in the development and functioning of all living organisms. Along with RNA and proteins, DNA is one of the three major macromolecules that are essential for all forms of life. Genetic information is encoded as a sequence of nucleotides (guanine, adenine, thymine, and cytosine) recorded using the letters G, A, T, and C. Most DNA molecules are double-stranded helices, consisting of two long polymers of simple units called nucleotides, molecules with backbones made of alternating sugars(deoxyribose) and phosphate groups, with the nucleobases (G, A, T, C) attached to the sugars. DNA is well suited for biological information storage, since the DNA backbone is resistant to cleavage and the double-stranded structure provides the molecule with a built-in duplicate of the encoded information. These two strands run in opposite directions to each other and are therefore anti-parallel, one backbone being 3' (three prime) and the other 5' (five prime). This refers to the direction of 3rd and 5th carbon on the sugar molecule is facing. Attached to each sugar is one of four types of molecules called nucleobases (informally, bases). It is the sequence of these four nucleobases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by copying stretches of DNA into the related nucleic acid RNA in a process called transcription. Within cells; DNA is organized into long structures called chromosomes. In the process of cell division, these chromosomes are duplicated in the process of DNA replication which provides each cell its own complete set of chromosomes. [3, 15, 16]

DNA repair refers to a collection of processes by which a cell identifies and corrects damage to the DNA molecules that encode its genome. Generally human cells, both normal metabolic activities and environmental factors such as UV light and radiation can produce DNA damage, resulting in as many as million individual molecular lesions per cell per day. Most of these lesions can cause structural damage to the DNA molecule and can alter or eliminate the cell's ability to transcribe the gene that the affected DNA encodes values. Other lesions induce potentially harmful mutations in the cell's genome, which affects the survival of its inherited cells after it undergoes mitosis. Subsequently, the DNA repair process is constantly active as it responds to damage in the DNA structure. Whenever normal repair processes getting failed, irreparable DNA damage may occur, and cellular apoptosis does not occur, including double-strand breaks and DNA cross linkages. The DNA repair ability of a cell is vital to the integrity of its genome and thus to its normal functioning and that of the organism. Many genes are initially shown to influence life span have turned out to be involved in DNA damage repair and protection. The Failure of correct molecular lesions in cells that form gametes can introduce mutations into the genomes of the offspring and thus influence the rate of evolution. The rate of DNA repair is

dependent on many factors which includes the cell type, the age of the cell, and the extra cellular environment.

2. ALGORITHM & TOOL

Decision Trees, logistics and Multilayer perceptron Algorithms are data-mining based classification methods for systematically analyzing data about human DNA repair genes. A linearly combined kernel with various classifier algorithms analyzes ageing and non-ageing data. The most familiar supervised learning algorithm enables better discrimination between ageing-related and non-ageing-related DNA repair genes. Classifiers algorithms allow better classification accuracy in DNA repair gene data set with linear combination of linear kernel and polynomial kernel of degree 3. Compared to Decision Trees and Multilayer perceptron, logistics Algorithm, classifiers with the proposed kernel can achieve 68% AUC (Area under ROC Curve) values, in contrast to 57.1% and 59.1% respectively. Classifier algorithm is a powerful and robust classification algorithm that can yield higher predictive accuracy values. Selection of proper kernel plays a more important role in fulfilling the classification task. This important genes identified not only can target critical pathways related to ageing but also detected genes that may reveal possible related ageing and non-ageing.

2.1 Multilayer Perceptron

A multilayer perception (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that is not linearly separable. [27]

2.2 Pruning Tree Algorithm

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The dual goal of pruning is reduced complexity of the final classifier as well as better predictive accuracy by the reduction of over fitting and removal of sections of a classifier that may be based on noisy or erroneous data. [12]

2.3 The Logistics Algorithm

Linear regression performs a least-squares fit of a parameter vector β to a numeric target variable to form a model

$$F(x) = \beta T \cdot x,$$

Where x is the input vector (Assume that the constant term in the input vector to accommodate the intercept). It is possible to use this technique for classified by directly match linear regression models to class indicator variables. If there are J classes then J indicator variables are created and the indicator for class j takes on value 1 whenever class j is present and value 0 otherwise. However, this approach is difficult from masking problems in the multiclass setting.

2.4 WEKA Data mining Tool

Open source libraries have also become very popular since the 1990s. The most prominent example is Waikato Environment for Knowledge Analysis (WEKA). WEKA started in 1994 as a C++ library, with its first public release in 1996. In 1999, it was completely rebuilt as a JAVA package. Since that time, it has been frequently updated. In addition, WEKA components have been integrated. WEKA toolkit is a widely used toolkit for machine learning and data mining that was originally developed at the University of Waikato in New Zealand. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License. More than twelve years have elapsed since the first public release of WEKA. In that time, the software has been rewritten entirely from scratch, evolved substantially and now accompanies a text on data mining. These days, WEKA enjoys widespread acceptance in both academia and business, has an active community, and has been downloaded more than 1.4 million times since being placed on Source Forge in April 2000. [26]

3. FEASIBILITY STUDY

DNA repair genes have been considered to recognize their properties which help in classifying them as ageing and non-ageing genes. Despite continuous efforts that have been put into analyzing the characteristics of these genes to classify them efficiently and good accuracy the adoption of a classifier like multilayer perceptron, logistics and pruning tree which is insensitive to these discrepancies would be ideal. A data mining approach is implemented for classifying DNA repair genes using various characteristics features. Data are present in genes data sets. The classification models built were difficult to analyze and dimensionality present in the gene data sets.

The main goal of this paper is to classify ageing and non-ageing genes. The following features are covered for classifications. The classification of DNA repair gens can be analyzed using data mining approach. Data are present in gens data sets. Data sets are a collection of data usually represented in tabular column. Each column represents a particular variable. Data sets comprise collection of data corresponding to the number of rows. The classification of genes forms a cluster. Then cluster of gene related data sets are evaluated with different algorithm.

The built classification models were difficult to evaluate their performance and deduce due to the annoyance of dimensionality present in the gene data sets. This problem could be solved by implementing Dimensionality Reduction which is a well-known preprocessing technique and it is reduce the data sets complexity to maintaining the integrity of original data sets. The feature compartment assortment technique along with various explore method is used to shrink the data set without changing the reliability of the original data sets the reduction in the data sets enabled the use of Multilayer perception and logistics in the efficient analysis of the data sets. Reduced data sets performance and original data sets performance could be evaluated by implementing various classifiers.

In gene expression pattern, to found that ageing genes tend to have higher co-expression coefficients with other genes than that of non-ageing genes in the gene expression profile. The ageing of the worldwide population means there is a growing need for research on the biology of non ageing and ageing. DNA damage is a key provider to the ageing process and elucidating the role of different DNA repair systems in ageing

is of great interest. In this paper propose a data mining approach, based on classification methods for analyzing data about human DNA repair genes and non repair genes. The goal is to build classification models that allow us to discriminate between ageing-related and non-ageing-related DNA repair genes, in order to well again recognize their different properties. These patterns and their analysis support non-homologous end joining double strand break repair as central to the ageing-relatedness of DNA repair genes. This work used for protein interaction partners to improve accuracy in data mining methods and this approach could be applied to other ageing-related pathways. The presence of missing values in data sets can affect the performance of a classifier constructed by those data sets as a training sample. A number of methods have been proposed to treat missing data and the one used more frequently is deleting instances containing at least one missing value of a feature. Feature subset selection is one of data preprocessing, which is of immense importance in the field of data mining. Feature subset selection step of data preprocessing approach with genetic algorithm (GA) and correlation based feature selection has been used in a cascaded fashion. Experiments have been carried out on medical data set which is publicly available at UCI.

4. Methodology

4.1 Gene Data Collections

Collection of gene related data sets should be maintained in single database and it's contained all relevant information of attributes about DNA Gene repair data sets. Data collection is any process of preparing, collecting and storing data, as part of a process improvement or similar research. The purpose of data collection is to obtain information of gene related data sets record, to make decisions about important issues, or to pass information on to other repositories. Data are preliminarily collected to provide information regarding a specific gene related attributes information.

4.2 Data Preprocessing

Data preprocessing describes any type of processing performed on DNA Gene Data sets to prepare it for another processing used as a fundamental data mining practice. It transforms the data into a unique format that will be more easily and effectively. It generally involves data cleaning, data integration, data transformation and data reduction and it is mainly concentrates on data cleaning and data reduction, as the data sets created is complex and contains missing values one of the well-known methodologies. Data preprocessing describes any type of processing performed on gene related data to prepare it for another processing procedure that will be more easily and effectively processed for solution.

4.3 Data Cluster Classification

Clustering is a process of partitioning a set of data set in a set of meaningful sub-classes and collection of data objects that are similar to one another and that can be treated as one group. In this module various algorithm is applied to DNA Gene repair Data sets and it is used to classifying data sets in accordance with ageing and non-ageing. Data clustering is a methodology in which, the information that is logically same gene related data is physically stored together. In order to increase the efficiency of search and the retrieval in database repositories, the number of memory accesses is to be minimized. In clustering objects are similar properties which are placed in one class of objects, a single access to the disk can retrieve the entire class. If the clustering takes place in some abstract algorithmic space, may group a population into subsets with similar characteristic, reduce the problem space by acting on only a representative from each subset and it is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters.

4.4 Clustered Data Comparison

Gene related data sets accuracy must be evaluated with different kinds of algorithm and each algorithm contained data sets result should be evaluated in accordance with accuracy and error rate ratio. Logistic regression is a generalization of linear regression. It is used primarily for predicting a performance of Gene related data sets. In producing the logistic regression equation, the maximum-likelihood ratio was used to determine the statistical significance of the ageing and non ageing variable. Logistic regression has proven to be very robust in a number of medical domains and is an effective way of estimating probabilities from gene variables. Multi layer perceptron and pruning process applied clustered gene should be analyzed with various attributes results like number of correctly classified data, number of incorrectly classified data, accuracy result of each algorithm and time and every algorithm result of gene data sets could be determined from accuracy and time consumption.

4.5 Data Visualization

Finally, comparison of DNA Gene Data set results will be visualized in graph format and it is easy to implement custom visualization which is performed by extracting relevant knowledge from DNA Gene data. Ageing and non ageing of gene data set should be represented as graph and also different algorithm processed data set result must be estimated and comparison performance is represented or visualized as graph or bar chart.

5. Results & Discussion

Implementation is the stage in the research where the theoretical design is turned into a working system and is giving confidence on the new system for the users, which will work effectively, efficiently, involves careful planning, investigation of the current classification system, its constraints on implementation, design of methods. The more complex implementation needs more system analysis and design effort required. It represents to the overall classifications of the genes and the ways in which that classification provides conceptual integrity for an analysis. In a broader sense however components can be generalized to represent major gene elements and their interaction.

The statistical method to analyze the topological features of genes in order to classify them into ageing and non-ageing which is attribute selection techniques can be used to classify very large biological data sets. The attribute selection method along with classifiers method and have used the reduced data sets for comparing the accuracies of various classification algorithm. Dimensionality reduction techniques have shown an increase in the accuracy of the models without losing the integrity of the original data sets. An integrated of a hybrid variety of the attribute selection methods can be used to classify data sets. It develops a model that can accurately label a novel gene. Important features of DNA repair genes to understand their unique characteristics which differentiate them from non-ageing repair genes. To design a model which efficiently and accurately classifies the data objects a number

of classification techniques can be used, by which analysis can be performed more easily when compared to the trial and error methods of in-vitro techniques. These techniques are very advantages in terms of their accuracy, cost of computation, scalability, simplicity and interpretability which makes the process of classification a less complicated one. Although these algorithms give accurate results, efficiency is an issue of concern in biological data mining due to their vast and complex data sets. The gene data set can be classified using different classification algorithm with the help of weka tool. The classification instances are represented in following graph which shows which clustered applied data sets of all three algorithm results should be evaluated in according to accuracy like times, classified correct data sets etc…

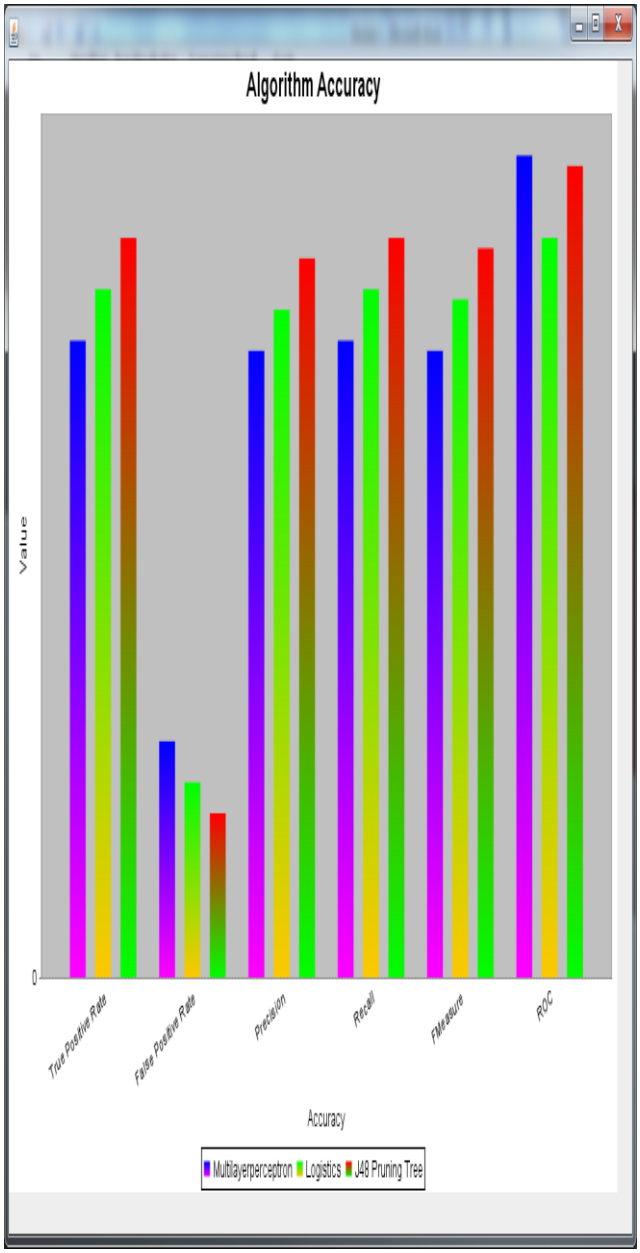

Fig 1: Comparison Graph

The following figure which clustered applied data sets of all three algorithm results should be evaluated in according to accuracy like times, classified correct data sets and it is shown in graph format.

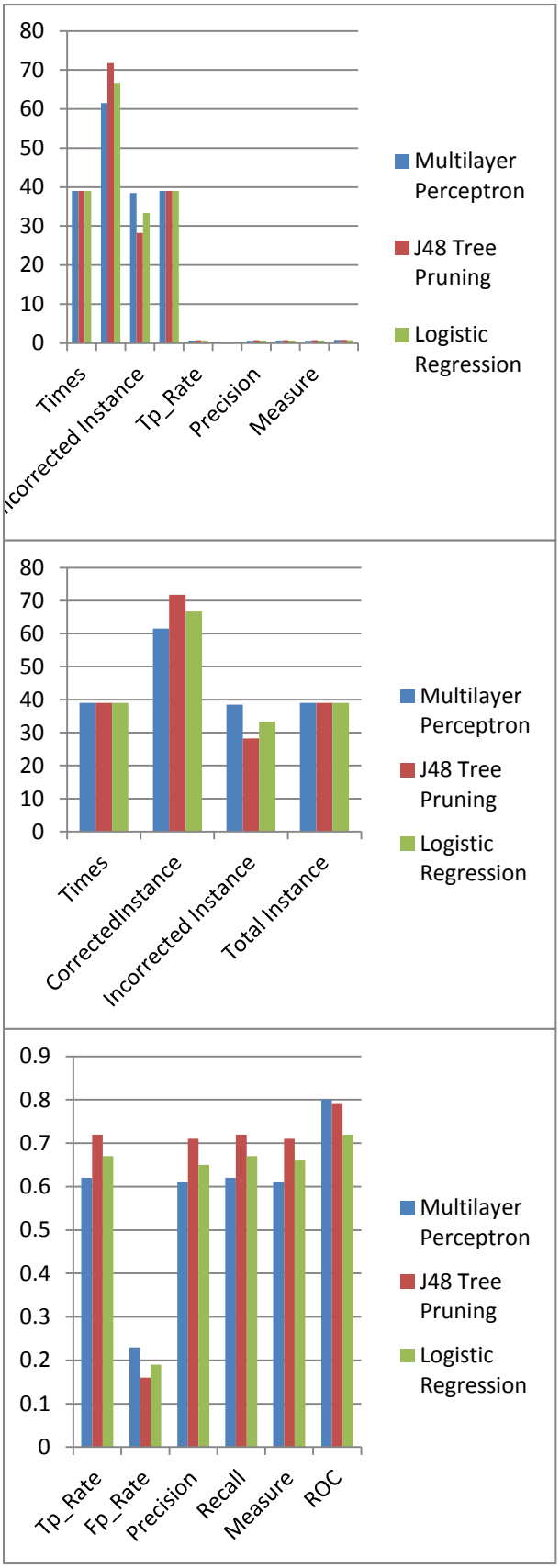

Fig 2: Comparison Chart

6. Conclusions & Future Work

The deployed data mining approach is used for classifying DNA repair genes using classification algorithm such as multilayer perceptron, logistics and pruning for analyzing the gene data about human DNA repair genes into ageing and non-ageing. The build classification models allow discriminating between ageing-related and non-ageing-related DNA repair genes, to understand their different properties and genes classification performance should be evaluated in according to ageing and non ageing related genes which are performed by applying different kinds of algorithm. This classification analyzes the characteristics of the genes to classify efficiently and accuracy the adoption of a classifier. This Classification on reduced data sets proves to be more efficient and interpretable making data analysis. Reduced data sets and dimensionality reduction is model friendly which helps the development of simpler and efficient classification models for classifying ageing and non ageing genes. This greatly contributes to the study of many complex diseases related to ageing.

Future work is to create a general framework for non ageing and ageing genes, simulate the effectiveness of gene oriented data sets should be classified by using data mining oriented effective algorithm. This result could be evaluated in terms of more accuracy. By using algorithm in the data set that result should be more efficient and less time consumption and quick retrieval of data. Generally for extracting the data from the data repository system take more time consumption and less accuracy. When using the data mining algorithm reduce the time consumption and provide more accuracy.

7. References

[1] Li. Y-H., Zhang, G. g.,Guo Z. "Computational Prediction of ageing genes in Human" In: Biomedical Engineering and Computer Science(ICBECS) 2010.

[2] Kenyon G. Chang J. Ganseh E. Rudner, A. Tabtiang R. " Mutant that lives twice as long as Wild Type" Nature, 1993.

[3] Freitas, A.A., Vasieva, (). Magalhaes, J. "A Data Mining Approach for classify DNA Repair Genes into Ageing Related or Non-Ageing Related" BMC Genomics- 2011.

[4] Han, J., Kamber, M. "Data mining concepts and ` Techniques" 2nd edn. Morgan Kauffmann 2006.

[5] Hall, M., Frank, E. "Combining Naïve Bayes and Decision Tables" In: 21st International FLAIRS - 2008.

[6] Asha Gowda, K., Jayaram. M.A., Manjunath, A.S "Feature Subset Selection using Cascaded GA and CFS: A Filter Approach in Supervised learning" International Journal of Computer Applications- 2011.

[7] Vasantha, M., Subbiah Bharathy, V. "Evaluation of Attribute Selection Methods with Tree Based Supervised Classification" International Journal of Computer Applications - 2010.

[8] Przulj N. Wigle A.D. Judicial I. "Functional Topology in a Network of Protein Interaction" Bioinformatics - 2004.

[9] Senescence. infor/genes/DNA-repair.html.

[10] Acuna, E., Rodreiguez c. "The Treatment of Missing Values and its Effect in the classifier Accuracy" In: Classification, Clustering and DM Application. - 2004.

[11] Tan P.N. Steinbach. M. Kumar V. "Introduction to Data Mining" 3rd edu. Pearson Education 2009.

[12] J. Zhang, V. Honavar, "Learning Decision Tree Classifiers from Attribute Value Taxonomies and Partially Specified Data", The 20th International Conference on Machine Learning, 2003.

[13] Harrison J.H. "Introduction to the Mining of Clinical Data. Clinics in Laboratory Medicine" Clinical Data Mining and Warehousing Volume 28, Issue 1 March 2008.

[14] D. Kang, K. Sohn, "Learning decision trees with taxonomy of propositional zed attributes", ACM Pattern Recognition, Volume 42, Issue 1, January 2009.

[15] Cohen A.M., Hersh W.R. "A Survey of Current Work in Biomedical Text Mining" Briefings in Bioinformatics.

[16] Gene Ontology: tool for the unification of biology. "The Gene Ontology Consortium" Nature Genet. 2000.

[17] B.Madasamy, Dr. J. Jebamalar Tamilselvi, "Optimal Data mining Classification Algorithm for Bio Medicinal Facts" International Journal of Advanced Computing Engineering Application (IJACEA) February -2013.

[18] B.Madasamy, Dr. J. Jebamalar Tamilselvi,"General Web Knowledge Mining Framework" International Journal of Computer Science and Engineering (IJCSE) October -2012.

[19] B.Madasamy, Dr. J. Jebamalar Tamilselvi,"Assessment of Freeware Data Mining Tools over some Wide Range Characteristics", SPRINGER Verlag Journal, ICIP - August 2012.

[20] B.Madasamy, Dr. J. Jebamalar Tamilselvi, "Wide Range Data Mining Application Framework using IT Decision Making" International Conference on Recent Trends in Computing Technology April 2013.

[21] George Hripcsak, Suzanne Bakken, Peter D. Stetson, and mla L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery" Journal of Biomedical Informatics 2003.

[22] Neeraj Kumar , Govind Kumar Jha "A Time Series ANN Approach for Weather Forecasting" International Journal of Control Theory and Computer Modeling (IJCTCM) January 2013

[23] Sinnott R.O. Stell A.J. Ajayi, O. "Supporting grid-based clinical trials in Scotland" Health Informatics Journal.

[24] Lalitha Saroja Thota, Suresh Babu Changalasetty "Optimum Learning Rate for classification Problem with MLP Data Mining" International Journal of Advances in Engineering & Technology, March. 2013.

[25] Brown D. "Introduction to Data Mining for Medical Informatics: Clinics in Laboratory Medicine", Clinical Data Mining and Warehousing, March 2008.

[26] http://www.cs.waikato.ac.nz/ml/weka/

[27] www.cs.waikato.ac.nz/~eibe/pubs