Survey on Medical Data Cluster analysis using Feature Selection and Neural Networks

V.Sangeetha, J.Preethi, M.Sreeshakthy

Abstract: **Medical data are the real world data and certainly complex and huge in number. Their analysis requires many complex operations. Such medical data are of image, signal or dataset. Works on classification of medical data are numerous whereas clustering medical data is on the rising side. In this paper, we make a survey on the procedures and its variants for making analysis on the medical image data using clustering. The images are preprocessed, features are extracted, then extracted features are selected using existing optimization algorithms and at last the features are clustered for analysis. To make clustering Neural Networks are discussed. The cluster is then validated for its accuracy.**

Keywords: **Medical data, Feature selection, Feature Extraction, Filter and Wrapper model, Clustering, Neural Networks, Cluster validation.**

I. INTRODUCTION:

Medical images are data with high dimensionality. In the recent years medical data are a variety of type. They are in image form, dataset form, signal form, wavelet form. With regard with datasets, they are prior aligned and require preprocessing in terms of their attributes. In case of signal form, their distortions must be removed. There may be many fluctuations and redundant noises in these signals. But at certain times, these noises must be checked before their removal because they may be of use. The images require lot of preprocessing for further analysis.

Preprocessing is the process of processing an image so as to remove the noise and outliers. There are a lot of preprocessing techniques for images. The result of preprocessing would be an image with less or no noise. Such image would yield perfect result when used for analysis. Once the image is preprocessed, they are used for feature extraction.Again there are many feature extraction techniques. Features are the dimensions or attributes of the images. Generally feature extraction requires domain knowledge. But at times, the general dimensionality reduction techniques may be of use. Once the features are extracted, they must be kept for optimal selection. Many search strategies are available that are to be discussed further through which necessary features are selected. Recently, biologically inspired optimization techniques are used for searching the feature search space. With the help of selected features, the process of clustering is performed. There are many algorithms and techniques available for supervised learning with feature selection. But algorithms for unsupervised learning with feature selection are in the developing side. This is so because of the absence of class label. Henceforth we use optimization at its best to get best accurate results. The resulting clusters are then validated further for its accuracy. The rest of the paper focuses on clustering with feature selection and neural networks. Section 2 discusses the Feature selection methods and their Related works. Section 3 discusses the Searching Strategies and a 3D framework. Section 4 discusses the abstracts of Unsupervised learning. Section 5 discusses Clustering, Neural Networks and their implementations in medical data. In the last, Section 5 concludes the concept of medical data analysis.

II. FEATURE SELECTION:

Feature selection also known as Attribute subset selection , Variable selection , Feature subset selection is the process of selecting subset or essential features from the existing set of features.

This process is done because the extracted features will be redundant and when this is used as a whole

***Manuscript received Nov 2014**.*

***V.Sangeetha**, Dept. of Computer Science, Anna University Regional Centre, Coimbatore-47, India.*
***J.Preethi**, Assistant Professor, Dept. of Computer Science, Anna University Regional Centre, Coimbatore-47, India.*
***M.Sreeshakthy**, Dept. of Computer Science, Anna University Regional Centre, Coimbatore-47, India.*

will reduce the accuracy of the process. The process of feature selection has certain stages. [1]

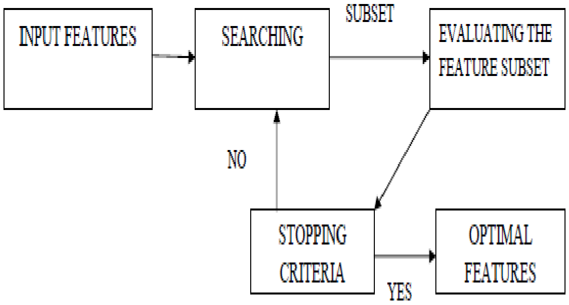


Fig 1:*Process of Feature Selection[1]*

The input features are given as such after eliminating certain discrepancies. The missing fields, repeated values, null values are certain kinds of discrepancies. Then the features are set for subset generation. Subset are generated through searching. There are three search strategies:

- Complete search
- Random search
- Sequential search

These search strategies are combined with the filter and wrapper models(to be discussed further) to create greater combinations. Once after searching, subsets are generated. Once the subsets are generated, the Attribute Evaluators are used to evaluate the subsets. Some of the evaluators are CfsSubsetEval [3], ChiSquaredAttributeEval [3], GainRatioAttributeEval [3], InfoGainAttributeEval [3]. These evaluators evaluate the quality of the subset generated. A Stopping criterion is used to stop the process subset generation.

Axiom-1:If there are N features, then there would be 2^N subsets of features.

The main problem in feature selection is the Curse of Dimensionality. In this regard, feature selection techniques are referred as Dimensionality reduction domain. They are of two types: Transformation based reduction and Selection based reduction. Feature selection would be the process of reducing the dimensionality of the dataset methodologically to produce optimal subsets of datasets. The main motive of feature selection is: 1) to reduce the over fitting of the data to the model [2] 2) to produce cost effective models[2]. The final result of feature selection would be however subset generation. The generated subsets are further evaluated for its saliency and relevance. The feature selection methods are categorized as – Filter and Wrapper methods.

A. FILTER METHODS:

The filter methods of feature selection is independent of any criteria or algorithm. These methods mainly use distance measures, consistency measures, information measures to select relevant features[1]. John et al [5] presented a definition for relevant features. F_iϵ F be a feature, S_iϵ S be a subset. Let s_i and f_i be the value assignment.

Definition 1[5]: F_i is said to be a irrelevant feature to the target if and only if there exists some s_i ,f_i and target for the probability $P(S_i-s_i\ ,\ F_i-fi)>0$.This indicates that when even if after a feature is removed from a set the probability of relevancy is greater, then the feature is irrelevant. Otherwise F_i is relevant.

B. WRAPPER METHODS:

The wrapper method of feature selection uses a predefined induction algorithm along with a search method. The subset evaluation is wrapped by a mining algorithm. Based on the mining algorithm accuracy, subsets are evaluated. These algorithms focus on Predictive accuracy and claim its best accuracy; however its computational cost will be more than filter method. These models highly support the use of randomized search for cancer research . But their drawback is their higher risk of over fitting of the model.

C. HYBRID METHODS:

Hybrid methods are the combination of both Filter and Wrapper methods. The predictive accuracy of wrapper methods and the low computational cost of filter methods are combined together to yield best results. Many research works have been performed

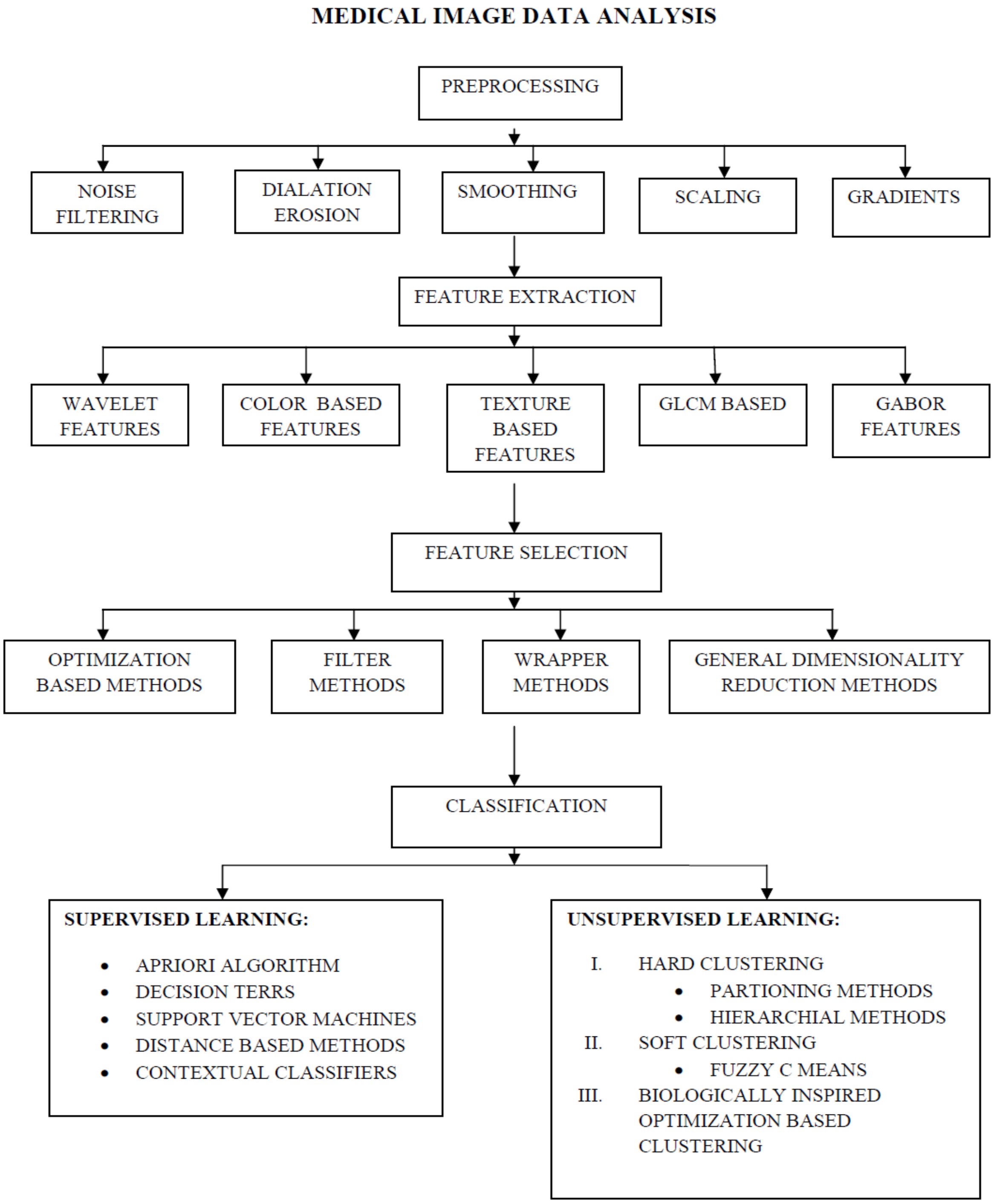
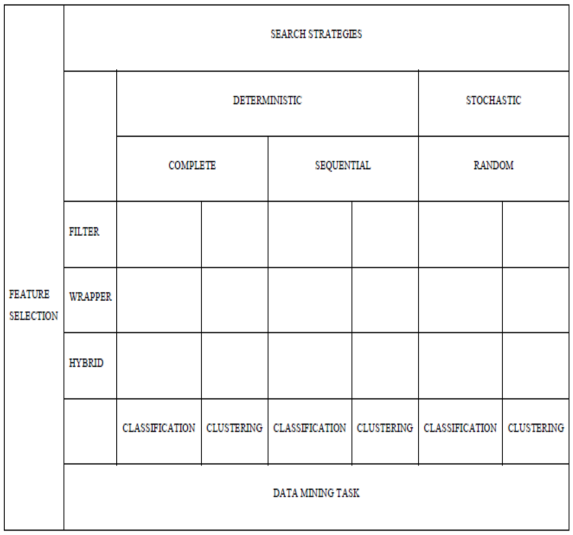


Fig 2: *Various methods for medical data analysis in procedural format*

III. SEARCH STRATEGIES:

Searching is of prime importance in feature selection. Searching can be either deterministic or stochastic. Deterministic search strategies are those strategies that do not support randomicity. They follow a predefined scheme. Some are Sequential search, Sequential forward search. The next category is the stochastic search. This uses the concept of randomness. All the evolutionary algorithms are in this category. They do not have a predefined strategy. This method is apt for high dimensional data. The Genetic algorithms, Optimization methods like PSO, Ant colony Optimization all fall under the stochastic search method. In [1], a three dimensional framework have been devised for all these strategies.

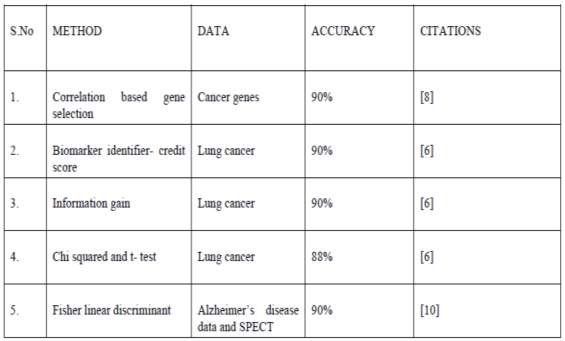


FEATURE SELECTION	SEARCH STRATEGIES						
		DETERMINISTIC				STOCHASTIC	
		COMPLETE		SEQUENTIAL		RANDOM	
	FILTER						
	WRAPPER						
	HYBRID						
		CLASSIFICATION	CLUSTERING	CLASSIFICATION	CLUSTERING	CLASSIFICATION	CLUSTERING
	DATA MINING TASK						

Fig 3: *A Three Dimensional Framework For Integrating Feature Selection, Searching and Data Mining Task[1]*

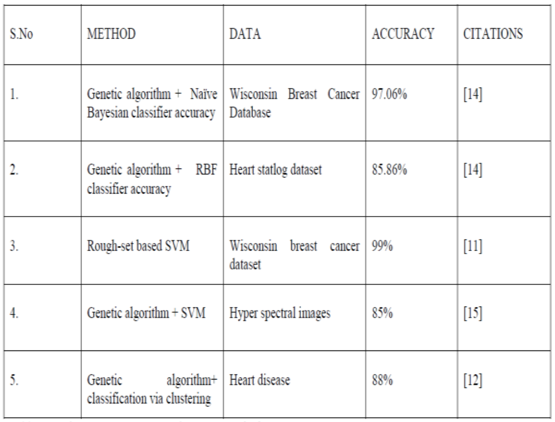
This three dimensional framework provides space for deriving new algorithms in the near future. Recently it is found that for the unsupervised learning under random search; algorithms are of interest and in need.

The following three tables shows the performance comparisons of various feature selection methods.



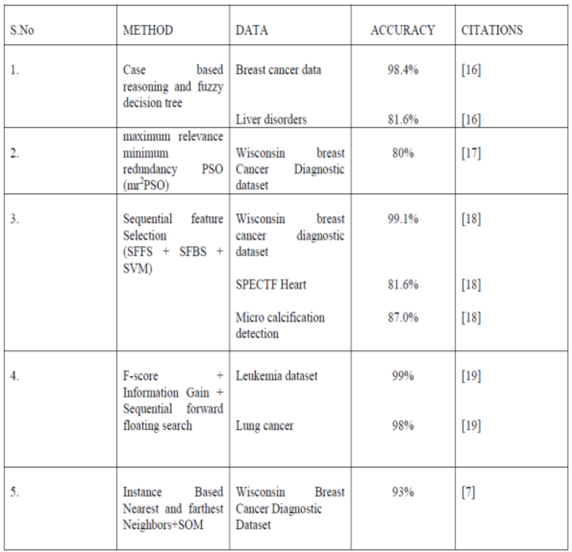
S.No	METHOD	DATA	ACCURACY	CITATIONS
1.	Correlation based gene selection	Cancer genes	90%	[8]
2.	Biomarker identifier- credit score	Lung cancer	90%	[6]
3.	Information gain	Lung cancer	90%	[6]
4.	Chi squared and t- test	Lung cancer	88%	[6]
5.	Fisher linear discriminant	Alzheimer's disease data and SPECT	90%	[10]

Table 1: *Performance Comparison of Filter Methods*



S.No	METHOD	DATA	ACCURACY	CITATIONS
1.	Genetic algorithm + Naïve Bayesian classifier accuracy	Wisconsin Breast Cancer Database	97.06%	[14]
2.	Genetic algorithm + RBF classifier accuracy	Heart statlog dataset	85.86%	[14]
3.	Rough-set based SVM	Wisconsin breast cancer dataset	99%	[11]
4.	Genetic algorithm + SVM	Hyper spectral images	85%	[15]
5.	Genetic algorithm+ classification via clustering	Heart disease	88%	[12]

Table 2: *Performance Comparison of Wrapper Methods*



S.No	METHOD	DATA	ACCURACY	CITATIONS
1.	Case based reasoning and fuzzy decision tree	Breast cancer data Liver disorders	98.4% 81.6%	[16] [16]
2.	maximum relevance minimum redundancy PSO (mr^{2}PSO)	Wisconsin breast Cancer Diagnostic dataset	80%	[17]
3.	Sequential feature Selection (SFFS + SFBS + SVM)	Wisconsin breast cancer diagnostic dataset SPECTF Heart Micro calcification detection	99.1% 81.6% 87.0%	[18] [18] [18]
4.	F-score + Information Gain + Sequential forward floating search	Leukemia dataset Lung cancer	99% 98%	[19] [19]
5.	Instance Based Nearest and farthest Neighbors+SOM	Wisconsin Breast Cancer Diagnostic Dataset	93%	[7]

Table 3: *Performance Comparison of Hybrid Methods*

IV. UNSUPERVISED LEARNING:

Unsupervised learning is so called because there would be no definite class information. The target information would be absent. Analysis in these cases would require special methods. Such situations are quite common in large database because assigning class labels for each data instance is difficult. Unsupervised learning is an example of *learning by observation* rather than learning by examples. Thus it has a greater impact in pattern recognition. Clustering is an example of unsupervised learning.

V. CLUSTERING:

Clustering is the process of grouping the data together into relevant groups without the prior knowledge of group definitions. The main objective of clustering is to transform the set of data into meaningful data so that data in the same group gives same sense. The main principle is to minimize the intracluster distance and maximize the intercluster distance [26]. Clustering is common in every aspect of life. It is also called data segmentation because of its ability to distinguish things. It also helps in outlier detection**.** Clustering has many applications like in pattern recognition, machine learning, business market analysis, detection of fraud in banks and forensic services. The main advantage of clustering is that it is adaptable to changes.

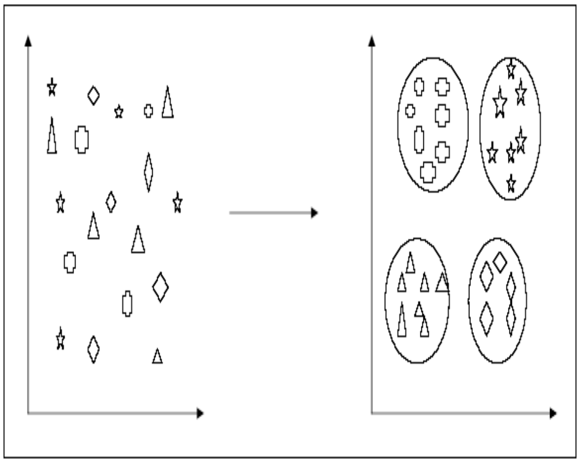


Fig 4: *Clustering process*

The commonly used clustering techniques are partioning clustering and hierarchical clustering. A Dendrogram[20] is usually used to represent the clustering tree. But since the advent of Neural Networks and Soft Computing techniques, clustering and their visualization is made much easier. Over the years, third generation networks are of interest.

A. CLUSTERING AND NEURAL NETWORKS:

NEURAL NETWORKS:

Neural networks are an implication of the real life neurons. The working patterns of the neural networks are similar to that of biological neurons. They get inspired when they are fired with the sense of stimuli. The input is encoded through the input neuron. Then the network is trained further for iterations. Many training algorithms are used like Genetic Algorithms based , Distance based, Hebbian learning[21], Back propogation learning [25], Perceptron learning[24],RBF, Linear Adaptive learning to train the network. The working of network is as follows:

Neurons are the basic processing unit of a neural network.

1. The inputs are passed on through the input neuron $\mathbf{x_{ij}}$.
2. The inputs are summed as in (2) and passed on to the hidden layer. The summation is performed using a linear combiner.
   $$u = \sum_{i=1}^{n} x_i w_{ij} + b \longrightarrow \quad (2)$$
3. Upon this an activation function is applied to limit the amplitude of the output neuron.
   $$y = \emptyset(u + b) \longrightarrow \quad (3)$$

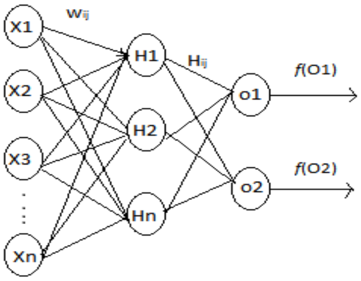


Fig 5: *An Artificial neural Network*

A neural network can be scaled to multiple layers. The accuracy of the networks is linearly dependent on the number of layers. The network can be trained either using supervised, unsupervised or reinforcement learning. The class definition decides the learning method.

NEURAL NETWORK AND ITS GENERATIONS:

With the advancements in sciences, neural network has also advanced greatly, from the basic model to the third generation Spiking neuron model.

First generation[21]: This uses the binary theory- occurrence and absence of spikes. This model is called the threshold gate and it is used in many multilayer perceptron networks, Boltzmann machine.

Second generation[21]: This model is called as sigmoidal gate. This model determines the firing rate of the spike(signal). The output is the number representing the firing rate. These are used in learning in the neural networks.

Third generation[21]: This is the so called Spiking neuron model. This generation uses the timing of spike firing. The output is the reference point that determines the time difference between spikes.

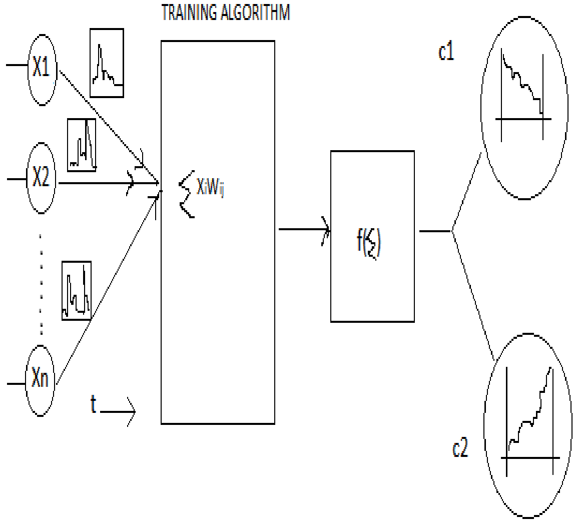


Fig 6: *A third generation Neural Network model*

B. NEURAL NETWORKS AND MEDICAL DATA CLUSTER ANALYSIS:

Neural networks and Kohonen map [22] are greatly applicable in medical data analysis. As the volume of data increases, data mining tasks play multispecialty roles. Applications of neural networks in this domain include tissue classification, micro calcification detection, image analysis, disease prediction, biochemical analysis and even in drug development. Artificial neural networks help doctors to great extent as they can easily process complex data. Neural networks can even works on the rules to make classification for disease prediction. Through this rule system it is easy to make prediction for a person's disease. In [23], Evolving topology based SNN is used make out classification of breast cancer data. Genetic Algorithm is used to train the network. Generally evolutionary algorithms are best in optimizing the training process. In [24], 3D Dielectrically heterogeneous breast cancer data are sort out and performance is compared against the existing LDA Classifier. Neural Network is the current trend for clustering and classification. In [25], heart disease is predicted using a single layer back propagation neural network. The network is trained using back propagation algorithm. In [27], image based clustering is done using Spiking Neural Networks. Brain images are used and the network is trained using Spike Prop Algorithm. It is feed-forward network. The brain images are taken their pixels are clustered as 8X8 matrix. The image is segmented using SNN. In [28], Wisconsin Breast Cancer Dataset is classified using a neural network. The network uses the Metaplasticity property of the neurons. The network is trained using the Artificial Metaplasticity Multilayer Perceptron algorithm.

C. PROS AND CONS:

Using neural networks greatly depends on the application domain. The dimensionality of the data decides the method. Neural networks are not self explanatory and they are less descriptive. They are purely analytical and moderately speedy. These networks are highly adaptable and scalable, thus they are fault tolerant. They can process even highly complex data.ANN are highly parallel and robust. They improve the performance through learning till the end. They are highly accurate. They show excellent performance in noisy environment. The aspects of clustering medical data, they are at most suitable.

VI. CONCLUSION:

Clustering medical data is purely unsupervised. It has no proper class definitions. For this clustering to be accurate, the features that are being used for must be salient and precise. Either of the filter or wrapper methods must be used to make out the property of feature saliency [1] and cluster compactness [20]. In this paper Neural Networks is being considered for clustering. The network must be trained to get the resultant clusters. Many training algorithms are depending on the type of data. The kind and size of network also depends on the type of the data. The resultant clusters are validated using DB Index [29][30] or R Square. The accuracy is studied from their graph. The future work for this survey can be extended to analyzing the individual feature selection algorithm with the generation of neural networks. Since feature selection for clustering are less in random search process, they are highly in the class for survey.

REFERENCES:

[1] H. Liu, L. Yu, Toward Integrating Feature Selection Algorithms For Classification And Clustering, IEEE Transactions On Knowledge And Data Engineering 17 (4)(2005) 491–502.

[2] G.Victo Sudha George And Dr. V.Cyril Raj, Review On Feature Selection Techniques And The Impact Of SVM For Cancer Classification Using Gene Expression Profile,

[3] Megha Aggarwal, Amrita, Performance Analysis Of Different Feature Selection Methods In Intrusion Detection, International Journal Of Scientific & Technology Research Volume 2, Issue 6, June 2013 ISSN 2277-8616

[4] Yubin Kuang, A Comparative Study On Feature Selection Methods And Their Applications In Causal inference, 26th May 2009.

[5] G.H. John, R. Kohavi, And K. P fleger, "Irrelevant Feature And The Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, Pp. 121-129, 1994.

[6] In-Hee Lee, Gerald H Lushington And Mahesh Visvanathan, A Filter-Based Feature Selection Approach For Identifying Potential Biomarkers For Lung Cancer, Journal Of Clinical Bioinformatics.

[7] Chien-Hsing Chen, A Hybrid Intelligent Model Of Analyzing Clinical Breast Cancer Data Using Clustering Techniques With Feature Selection, Elsevier, Applied Soft Computing 20 (2014) 4–14

[8] Yongjun Piao, Minghao Piao, Kiejung Park And Keun Ho Ryu , An Ensemble Correlation-Based Gene Selection Algorithm For Cancer Classification With Gene Expression Data, Bioinformatics, Vol. 28 No. 24 2012, Pages 3306–3315

[9] Jing Kong, Sijianwang And Gracewahba, Using Distance Covariance For Improved Variable Selection With Application To Genetic Risk Models,14077297 V3,2 Sep 2014

[10] J.Ramírez, J.M.Górriz , D.Salas-Gonzalez , A.Romero ,M.López, I.Álvarez , M.Gómez-Río, Computer-Aided Diagnosis Of Alzheimer's Type Dementia Combining Support Vector Machines And Discriminant Set Of features, Elsevier, Information Sciences 237 (2013) 59–72

[11] Hui-Ling Chen , Bo Yang , Jie Liu , Da-You Liu, A Support Vector Machine Classifier With Rough Set-Based Feature Selection For Breast Cancer Diagnosis , Elsevier, Expert Systems With Applications 38 (2011) 9014–9022

[12] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, Enhanced Prediction Of Heart Disease With Feature Subset Selection Using Genetic Algorithm, International Journal Of Engineering Science And Technology Vol. 2(10), 2010, 5370-5376

[13] J.G. Dy, C.E.B.A. Kak, L.S. Broderick, A.M. Aisen, Unsupervised Feature Selection Applied To Content-Based Retrieval Of Lung Images, IEEE Transactions On Pattern Analysis And Machine Intelligence 25 (3) (2003) 373–378.

[14] Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath , Feature Subset Selection Problem Using Wrapper Approach In Supervised Learning, 2010 International Journal Of Computer Applications (0975 – 8887),Volume 1 – No. 7 .

[15] Shijin Li , Hao Wu, Dingsheng Wan, Jiali Zhu, An Effective Feature Selection Method For Hyper spectral Image Classification Based On Genetic Algorithm And Support Vector Machine, Knowledge-Based Systems 24 (2011) 40–48.

[16] Chin-Yuan Fana, Pei-Chann Changb, Jyun-Jie Linb, J.C. Hsiehb , A Hybrid Model Combining Case-Based Reasoning And Fuzzy Decision Tree For Medical Data Classification, Elsevier, Applied Soft Computing 11 (2011) 632–644.

[17] Alper Unler , Alper Murat , Ratna Babu Chinnam, Mr2pso: A Maximum Relevance Minimum Redundancy Feature Selection Method Based On Swarm Intelligence For Support Vector Machine Classification, Elsevier, Information Sciences 181 (2011) 4625–4641.

[18] Yonghong Peng , Zhiqing Wu, Jianmin Jiang, A Novel Feature Selection Approach For Biomedical Data Classification, Elsevier, Journal Of Biomedical Informatics 43 (2010) 15–23.

[19] Hui-Huang Hsu, Cheng-Wei Hsieh , Ming-Da Lu, Hybrid Feature Selection By Combining Filters And Wrappers, Elsevier, Expert Systems With Applications 38 (2011) 8144–8150.

[20] Juha Vesanto And Esa Alhoniemi, Student Member, IEEE, Clustering Of The Self-Organizing Map, IEEE Transactions On Neural Networks, Vol. 11, No. 3, May 2000.

[21] Thomas Natschläger, December 1998, Networks Of Spiking Neurons: A New Generation Of Neural Network Models.

[22] P.Dostál ,P.Pokorný, Cluster Analysis And Neural Network.

[23] Edén A. Alanís-Reyes, José L. Hernández-Cruz, Jesús S. Cepeda, Camila Castro, Hugo Terashima-Marín, Santiago E. Conant-Pablos, Analysis Of Machine Learning Techniques Applied To The Classification Of Masses And Micro calcification Clusters In Breast Cancer Computer-Aided Detection, Journal Of Cancer Therapy, 2012, 3, 1020-1028.

[24] M. O'halloran, B. Mcginley, R. C. Conceicao, F. Morgan ,E. Jones And M. Glavin, Spiking Neural Networks For Breast Cancer Classification In A Dielectrically Heterogeneous Breast, Progress In Electromagnetics Research, Vol. 113, 413{428, 2011.

[25] Dr. K. Usha Rani, Analysis Of Heart Diseases

Dataset Using Neural Network Approach, International Journal Of Data Mining & Knowledge Management Process (Ijdkp) Vol.1, No.5, September 2011.

[26] P. Berkhin, A Survey Of Clustering Data Mining Techniques.

[27] Miss.Chinki Chandhok, Mrs.Soni Chaturvedi, Adaptation Of Spiking Neural Networks For Image Clustering, International Journal Of Video & Image Processing And Network Security Ijvipns-Ijens Vol: 12 No: 03.

[28] A. Marcano-Cedeño , J. Quintanilla-Domínguez, D. Andina, Wbcd Breast Cancer Database Classification Applying Artificial Metaplasticity Neural Network, Elsevier, Expert Systems With Applications 38 (2011) 9573–9579.

[29] K. Wang, B. Wang, L. Peng, Cvap: Validation For Cluster Analyses, Data Science Journal 8 (2009) 88–93.

[30] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On Clustering Validation Techniques, Journal Of Intelligent Information Systems 17 (2001) 107–145.