Handling Missing Values in Data Mining - A Case Study of Heart Failure Dataset

N. Poolsawad L. Moore C. Kambhampati and J. G. F. Cleland

Distributed Reliable Intelligent Systems Research Group (DRIS)
Department of Computer Science
University of Hull
Cottingham Road
Hull, HU6 7RX
United Kingdom
N.Poolsawad@2008.hull.ac.uk
Lisa.Moore@2011.hull.ac.uk
C.Kambhampati@hull.ac.uk

Hull York Medical School
Department of Cardiology
University of Hull
Cottingham Road
Hull, HU6 7RX
United Kingdom
J.G.Cleland@hull.ac.uk.

Abstract— In this paper, we investigate the characteristics of a clinical dataset using feature selection and classification techniques to deal with missing values and develop a method to quantify numerous complexities. The research aims to find features that have high effect on mortality time frame, and to design methodologies which will cope with the following challenges: missing values, high dimensionality, and the prediction problem. The experimental results will be extended to develop prediction model for HF This paper also provides a comprehensive evaluation of a set of diverse machine learning schemes for clinical datasets.

Keywords— data mining, feature selection, missing values, classification, clinical dataset, heart failure

I. INTRODUCTION

Data mining aims to automatically extract knowledge from large scale data; however information and knowledge mined must be meaningful enough to lead to some advantages. This method ties many technical areas, including machine learning, human-computer interaction, databases and statistical analysis. Clinical datasets have posed a unique challenge for data mining algorithms due to high dimensionality, multiple classes and values, noisy data, various systematic and human errors also presented with a number of features [1]. In addition, uncertainty of clinical data becomes widely available in the data mining techniques [2].

Currently large amounts of clinical data exist; however accurate models for predicting survivability of patients with heart failure are not extensively available. Identifying robust predictive models has proven to be a difficult task due to the nature of the availability of clinical data, in that there are large number of variables, a great deal of missing data, non-normal distributed data and unbalanced classes where one class is represented by a large number of samples and the other represented by a few numbers. The dataset used in this study is a large cardiological database called LIFELAB, a prospective cohort study consisting of 463 variables and 2,032 patients who were recruited from a community-based outpatient clinic based in England, the University of Hull Medical Centre, UK. Variables consisting of more than 20% missing values were disregarded, therefore substantially reducing the number of variables and patients to 20 variables and 1,051 patients. This implies that the data consisted of multiple missing values that either needed to be replaced or in all eliminated to allow appropriate analysis and algorithmic implementation.

Herein, the properties of various feature selection schemes are considered vis à vis the heart failure clinical datasets. The processes employed to predict models for designing treatments for patients with heart failure are in the following sequence; pre-processing, feature selection, classification, and evaluation. We strive to transform the dataset into an appropriate form so that data mining algorithms can be successively applied. As a conclusion we discuss the analyzed results and describe the focus on future research. A brief methodology regarding the data mining processes are discussed in the following topics:

II. DATA MINING PROCESSES

The goal of data mining in health care systems is to assist clinicians to improve the quality of prognosis and/or diagnosis and to facilitate the timelines of the medical problem. The target problems were extracted from the dataset using various data mining processes which is the prediction of the mortality and mobility time frame of patients with heart failure. The procedure follows a four step methodology. 1) Pre-processing the datasets to manually remove any redundant data and unnecessary variables which involve handling missing values, using normalization to ensure that the data elements are within the same scale in order to achieve both data integrity and performance. The process also involves four different missing value replacement methods namely; mean imputation, expectation-maximization (EM) algorithm, k-nearest neighbor (k-NN) imputation, and artificial neural network (ANN) imputation. Most datasets encountered in practice contains missing values and most machine learning schemes lack the ability to handle such datasets, in this study we have replaced missing values with imputation values. In addition, Sehgal *et al.,* [3] proposed an innovative missing value imputation algorithm known as Collateral Missing Value Estimation (CMVE) which uses multiple covariance-based imputation matrices for the final prediction of missing values. The authors also suggested that least square regression and linear

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

programming methods are used to optimize and compute the matrices [3]. 2) Three feature selection techniques; *t*-Test [4], entropy ranking [5, 6] and nonlinear gain analysis (NLGA) [15] are employed. These methods use a feature importance measure according to their individual discriminative capability to select the most relevant features, therefore reducing the size of the dataset, also evaluates classification performance measures while considering appropriate features.

Feature selection, also known as subset selection is a process that selects the most relevant attributes and attempts to find the best subset of the input feature set. Feature selection reduces system dimension, complexity and processing time while improving system performance on some dependent measure. Feature selection has two models: one is a wrapper model and the other is a filter model [7]. The wrapper model uses a predictive accuracy of a pre-determined learning algorithm to determine the goodness of the selected subset. The learning algorithm is run with various subsets of features and the learner that performs the best is chosen. In contrast, the filter model presents the data with the chosen subset of features to a learning algorithm. It separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm [8].

3) Classifier; multilayer perceptron (back-propagation), J48 (decision tree) and radial-basis function network (RBFN) are classification techniques were implemented. The results will reveal the performance of classification from the different techniques of missing value replacement methods, feature selection and classifier. 4) Two clustering techniques; K-means and hierarchical [9] are both explored to ascertain similarities between features in the dataset. *K*-means clustering assumes that the number of clusters, *k* is fixed [10, 11], the cluster is represented within cluster means of the variables and by a list of entities. The means in *k*-means signifies an aggregation of clusters and are usually referred to as centroids. We employed hierarchical clustering to reveal similarities and relationships between variables. The method partitions the data into a division of clusters and points, at each stage of the process the clusters are combined in a different layer of the hierarchy. This is visualized through the use of a dendrogram. In addition, two issues should be considered in practice; 1) deciding on the number of clusters to apply for each clustering algorithm and 2) defining the categorical attributes [9, 12]. In the present study, the number of cluster will be fixed to ensure a fair and consistent analysis and different categorical attribute are present in the dataset, each representing a different test. It is important to bear in mind that defining categorical data can be a difficult task in clustering analysis [13], for this reason the underlined clustering algorithms are implemented to achieve the best possible clustering outcome based on their respective theory.

III. EXPERIMENTAL RESULTS

The results presented here illustrate the method introduced in the data mining processes. Firstly, the data are pre- processed for further analysis and then explored to discover data characteristics. For the sake of precision and consistency, the meaningful and relevant variables were used for implementation.

TABLE I. THE STATISTIC OF VARIABLES MISSING VALUE HANDLING

Variable	Statistic	Missing Value Imputation				
		Original	EM	k-NN	Mean	ANN
Glucose	%missing	4.19				
	mean	0.088	0.088	0.088	0.088	0.089
	SD	0.060	0.059	0.059	0.059	0.060
	#data	886	925	924	929	933
Haemoglobin	%missing	0.95				
	mean	0.577	0.577	0.457	0.577	0.577
	SD	0.131	0.131	0.107	0.131	0.131
	#data	709	716	745	719	715
MCV	%missing	20.74				
	mean	0.795	0.795	0.811	0.795	0.788
	SD	0.066	0.061	0.068	0.059	0.063
	#data	706	892	830	900	897
Iron	%missing	13.51				
	mean	0.262	0.329	0.258	0.262	0.327
	SD	0.127	0.112	0.119	0.118	0.105
	#data	671	759	786	751	759
Vitamin B12	%missing	7.04				
	mean	0.094	0.094	0.094	0.094	0.093
	SD	0.062	0.060	0.060	0.060	0.068
	#data	863	925	927	929	955
Red Cell Folate	%missing	8.75				
	mean	0.229	0.231	0.229	0.229	0.073
	SD	0.141	0.137	0.135	0.135	0.046
	#data	767	840	840	842	937

TABLE II. DATA DISTRIBUTION OF DIFFERENT VARIABLES.

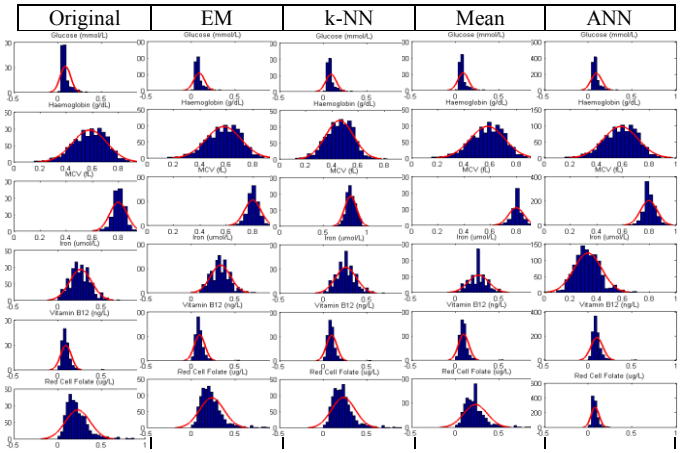

Table I shows the variables; with approximately 1-20% missing values. The table compares the statistical values between the original data and the data treated with different missing value imputation methods. Comparing the standard deviation (σ) and mean (μ) after handling the missing value by missing value imputation method, these values changed when compared with the original data variables consisting of missing values. The #data is the numbers of data points within the normal distribution range and these data points are within the range of [μ-σ, μ+σ]. The missing value replacement methods (EM, k-NN, Mean and ANN) show an increase in the number of data under the distribution curve. Table II shows the data

distribution of the different missing value imputation techniques. The techniques illustrate a differentially distributed result when compared with the same variables. For example in Table I and II k-NN produces the best results for Haemoglobin and Iron, while ANN shows the most accurate results for Glucose, Vitamin B12, and Red Cell Folate, and the mean imputation is suitable for MCV. However, the techniques for handling missing values gives different results due to their individual function and task.

Table III and IV shows the precision and recall values of the different missing value replacement and feature selection methods. We have shown the percentages of classification by classes of outcomes, Table III presents the outcome of mortality of patients in dead/alive classes, the precision and recall values that has the most precise values, meaning values approximately 100%, these are those treated using neural network techniques; MLP and RBFN. NLGA is the most effective method for decision tree, and the results in Table III and Table IV reflects its accuracy due to the consistent values. Table IV shows the outcome of the time periods of death which consist of 6 classes (6, 12, 18, 24, 36, >36). While Table III consists of two classes (dead/alive), as a result this indicates an accurate precision and recall values.

TABLE III
THE CLASSIFICATION RESULTS FROM DIFFERENT MISSING VALUE REPLACEMENT METHODS AND FEATURE SELECTION (FS) TECHNIQUES BY DEAD AND ALIVE CLASSES.

FS	Classifier	Missing values	EM Algorithm		k-NN imputation		Mean imputation		ANN imputation	
		Class	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
t-Test	MLP	Precision	81.9	81.8	76.1	82.1	81.6	81.4	77.8	82.8
		Recall	58.9	93.4	61.2	90.3	57.8	93.4	62.6	91
	DT	Precision	87.7	89.8	95.9	90.3	93.1	92.3	96.2	93.1
		Recall	78.8	94.4	79	98.3	84.1	96.8	85.6	98.3
	RBFN	Precision	100	96.81	99.7	96.94	100	96.81	100	96.81
		Recall	93.48	100	93.77	99.86	93.48	100	93.48	100
Entropy	MLP	Precision	72.5	78.6	70.5	78.8	71.1	77.9	71.3	79.3
		Recall	51.6	90.1	52.7	88.8	49.6	89.8	54.1	89
	DT	Precision	93.2	89.4	86.5	88.5	87.3	91	91.6	91.8
		Recall	77.3	97.1	75.9	94	81.6	94	83	96.1
	RBFN	Precision	99.7	97.48	100	98.31	99.7	97.76	99.7	97.76
		Recall	94.9	99.86	96.6	100	95.47	99.86	95.47	99.86
NLGA	MLP	Precision	77.5	80.3	77.2	80.7	74.6	79.9	76.5	77.3
		Recall	55.5	91.8	56.7	91.5	55	90.5	46.2	92.8
	DT	Precision	93.1	92.6	79.9	88.5	79.2	84.9	98	87.2
		Recall	84.7	96.8	76.8	90.3	68	91	71.1	99.3
	RBFN	Precision	100	97.08	100	97.08	100	97.76	99.7	97.35
		Recall	94.05	100	94.05	100	95.47	100	94.62	99.86

TABLE IV
THE CLASSIFICATION RESULTS FROM DIFFERENT TYPE OF MISSING VALUE REPLACEMENT METHODS AND FEATURE SELECTION TECHNIQUES BY MORTALITY CLASSES OF MONTHS.

Missing values			EM Algorithm						k-NN imputation					
Class (Months)			6	12	18	24	36	>36	6	12	18	24	36	>36
t-Test	MLP	Precision	76.5	61.9	83.3	42.6	34.6	49.6	73.6	59.7	55.6	44.2	70	49.1
		Recall	43.8	34.7	1.85	32.8	42.4	86.2	59.6	53.3	18.5	31.1	21.2	89.5
	DT	Precision	87.2	84	85.1	90.6	77.6	91.6	88.4	86.3	86.7	79.7	79.7	92.2
		Recall	84.3	90.7	74.1	78.7	89.4	92.8	85.4	92	72.2	83.6	83.3	92.8
Entropy	MLP	Precision	53.9	29.8	40.8	75	36	48.6	59.3	39.8	48.3	90	39	50.2
		Recall	46.1	37.3	37	9.8	13.6	78.3	53.9	44	25.9	14.8	34.8	77.6
	DT	Precision	88.6	85.2	86.4	82.5	86.2	87.7	87.9	87.2	84.9	82	79.7	93.1
		Recall	87.6	92	70.4	77	84.8	93.4	89.9	90.7	83.3	82	83.3	88.8
NLGA	MLP	Precision	71	42.2	51.7	50	30.9	57.4	55.3	49	52.6	100	33.9	46.3
		Recall	49.4	61.3	27.8	16.4	31.8	78.9	47.2	32	18.5	16.4	31.8	85.5
	DT	Precision	92.8	88	87.3	89.1	88.9	88	86.9	88.4	89.6	82.5	74	86.4
		Recall	86.5	88	88.9	80.3	84.8	96.1	82	81.3	79.6	77	86.4	92.1
Missing values			Mean imputation						ANN imputation					
Class (Months)			6	12	18	24	36	>36	6	12	18	24	36	>36
t-Test	MLP	Precision	57.3	41.9	55.6	55.6	29.1	59.3	82.6	60	62.5	54.2	40.7	54
		Recall	57.3	41.3	27.8	24.6	37.9	75.7	64	48	27.8	21.3	50	84.9
	DT	Precision	86.2	86.3	89.8	85.7	87.1	88.3	91.9	84.8	88.1	87.7	84.5	89.5
		Recall	91	84	81.5	78.7	81.8	94.7	88.8	89.3	68.5	82	90.9	95.4
Entropy	MLP	Precision	63.5	43.6	37.5	100	34.5	46.5	82	58.2	77.8	82.4	37.9	46.5
		Recall	52.8	22.7	27.8	8.2	28.8	86.8	56.2	42.7	25.9	23	33.3	88.
	DT	Precision	87.6	76.5	87.5	77.8	84.6	89.7	86.9	87.5	91.1	91.1	80.6	82.7
		Recall	87.6	86.7	64.8	80.3	83.3	91.4	82	84	75.9	83.6	81.8	94.1
NLGA	MLP	Precision	85.7	52.9	53.8	45	47.2	47.5	52.7	83.8	42.9	67.9	37.8	53.8
		Recall	47.2	36	25.9	29.5	37.9	86.8	66.3	41.3	22.2	31.1	47	74.3
	DT	Precision	86.7	84	86.3	87	87	92.8	96	87.3	90.9	79.7	85.3	84
		Recall	87.6	90.7	81.5	77	90.9	92.8	80.9	82.7	74.1	83.6	87.9	96.7

TABLE V THE SELECTED FEATURES

No.	Outcome	
	Mortality (Dead/Alive)	Mortality time frame (Dead Month)
1	Potassium	Sodium
2	Chloride	Bicarbonate
3	Urea	Urea
4	Creatinine	Creatinine
5	Calcium	MR-proANP
6	Phosphate	CT-proAVP
7	Bilirubin	Haemoglobin
8	Alkaline Phophatase	White Cell Count
9	ALT	Platelets
10	Total Protein	Total Protein
11	Albumin	Bilirubin
12	Triglycerides	Alkaline Phophatase
13	Haemoglobin	Adj Calcium
14	Iron	Phosphate
15	Vitamin B12	Cholesterol
16	Ferritin	Uric Acid
17	TSH	CT-proET1
18	MR-proANP	Red Cell Folate
19	CT-proET1	Ferritin
20	CT-proAVP	NT-proBNP

TABLE VI: INDICATES CORRELATION COMPARISON

Test variables	Correlation	Similarity levels
Creatinine and Urea	0.8	90.7
MR-proANP and CT-proET1	0.6	79.9

Table V shows the features selected (in bold) using the ANN imputation and NLGA feature selection technique. The result compares the selected features in both outcomes (Mortality-dead/alive and mortality time frame (Dead Month) and it indicates the variables highlighted appeared in both outcomes. This signifies that both applied techniques are capable of locating significant variables in the dataset.

Fig. 1 shows three clusters of two distinctive dead and alive classes, alive patients are indicated in red and dead patients are represented in black, the green cluster is suggested to be patients predicted to be alive with a few projected towards the dead group. While Fig. 2 below illustrates four clusters grouped into two classes of dead and alive, with the blue cluster represented as dead patients.

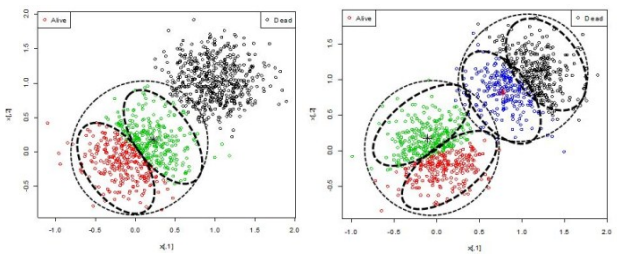

FIGURE 1. K-MEANS CLUSTERING INDICATING THREE CLUSTERS.
FIGURE 2. FOUR CLUSTERS OF THE DATA ARE ILLUSTRATED.

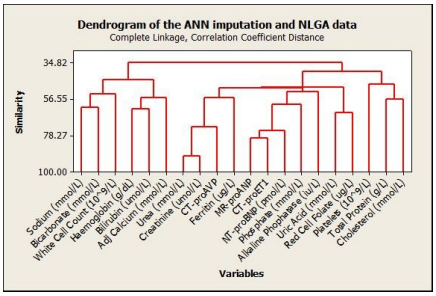

FIGURE 3. DENDROGRAM USED IN HIERARCHICAL CLUSTERING

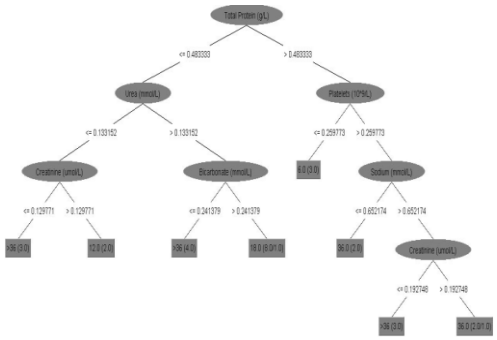

FIGURE 4: DECISION TREE FOR PREDICTING THE MORTALITY MONTHS

Fig. 3 demonstrates the relationship and similarities between the variables and as indicated by the dendrogram, urea and cteatinine are the most similar followed by MR-proANP and CT-proET1. This signifies a clear relationship between the variables and correlation values shown in Table VI which further supports their relation and similarity.

The selected features shown in Table V were utilized to construct the decision tree (Fig. 4); neural network was used for filling in missing data and NLGA for selecting feature as it shows the highest percentage of classification measurement.

IV. DISCUSSION

This paper has revealed that handling missing values using the underlined techniques is significant in data mining processes. Missing value is the major issue as it affects the feature reduction and classification processes and as a result the missing value imputation method was found to solve this problem. The results in Table I and II shows the different number of data points appeared under the normal curve in the different imputation techniques applied and those methods that illustrate the maximum number of data under the curve are suggested to be suitable and appropriate for data mining process. The mean and k-NN imputation techniques used the μ of the data variable to replace the missing values while k-NN calculates the μ of k nearest variables. As a result, it can be concluded that both methods produced similar results. The EM algorithm estimates values by using maximum likelihood. The EM algorithm results shown in Table I and II fails to illustrate the highest number of data (Table I) under the distribution curve (Table II), as the method has its limitations towards normally distributed data [1]. In contrast, the ANN imputation shows an increase in the number of data under the distribution curve. In addition, the missing value imputation techniques has shown to maintain the size of the datasets and also allows one to use all data types including categorical and numerical data, therefore the technique is useful for handling missing values in large dataset. Further, handling missing data with an inappropriate algorithm or technique can lead to biased, invalid or insignificant results; hence it is vital to approach missing values with methods specific for that particular dataset.

Applying feature selection methods such as entropy, the percentage of measurement from EM algorithm (Table III and IV) to fill missing values were the maximum compared to other missing value replacement algorithms. In literature, it is known that the EM algorithm uses the Kullback-Leibler distance (KL) [14], also known as relative entropy, which defines a distance measure between probability distribution, similar to entropy ranking for feature selection. When considering the classification results the performance with dead/alive class (Table III) shows a better precision and recall results due to the number of classes. The dataset of mortality classes (Table IV) leads to imbalanced class and the distribution are not even; this poses the challenges in terms of classification accuracy. Comparing the number of classes between Table III and IV, table III is mortality which has two classes and table IV has six classes of mortality months. Table III shows the classification results of the three dimensionality reduction techniques i.e. feature selection algorithm (*t*-Test, entropy and NLGA). The variables selected using the *t*-Test reduction method are significantly useful to develop the model for predicting heart failure because it selected Triglycerides, Potassium, Urea/ Uric acid, Creatinine, Nt-proBNP, and sodium as strong associations with mortality of heart failure [16, 17]. The present finding shows the metrics of accuracy, precision, and recall, all indicate feature selection to be a sufficient method for improving classification accuracy. Reference [7] argues that in theory, more features should provide more power, but in practice only significant features are efficient and this is reflected in the experimental results section. In addition, the executed dimensionality reduction techniques (feature selection) in this study, selects a small subset from an entire set of features [18].

Feature selection has been successfully applied to clinical dataset relating to the following areas e.g., lymphoma, gene expression, cancer [4, 6, 19, 20]. Reference [18] claimed that feature selection consistently increases accuracy, reduced feature set size, and provided better accuracy of classification. Reference [6] states that feature selection played an important role in classification. Effective in enhancing learning efficiently, increasing productive accuracy and reducing complexity of learning results, in addition learning is efficiently achieved with just relevant and non-redundant features.

In theory, data would be precisely distributed but in the real world situation data distribution are usually imprecise. Feature selection depends on the nature and how the data is distributed. The pre-processing step provides the story behind the data and tends to understand the nature of data, therefore allows the opportunity to choose the appropriate feature selection technique. The mean and standard deviation describes the group of data. For example entropy is related with the data density and finds the maximum distance between the target classes.

The clustering algorithms employed in this study have shown that the dataset is structured in an unsupervised manner to simplify the process of information retrieval. This finding correlates with works by Bean and Kambhampati [21], both exploited this notion by representing knowledge extracted from real data in the form of a decision rule set with minimal ambiguity to support in decision making. This was accomplished by employing clustering analysis and rough set theory, also explored the conceptual differences and similarities as well as the link between the two techniques [22].

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we have investigated missing value imputation techniques for handling missing values and dimensionality reduction based on feature selection. We attempted to understand and find suitable techniques for developing the model for analysis in clinical dataset. The selected features during the feature selection process were chosen based on the optimal criteria. The effect of these complexity measures on classification accuracy were evaluated using three diverse machine learning algorithms: multilayer perceptron (back-propagation), J48 (decision tree), and RBFN (neural network). Herein we have concluded that feature selection is one of the most useful tools for developing the prediction model because the decision support system requires meaningful and significant feature to make a decision to create an effective model. From the experimental results, the feature selection and missing value handling methods has gained potential performance of classification for predictive modeling. The key factor is to understand the nature of the dataset in order to choose the suitable technique. The important outcomes of extensive study will help in choosing the suitable handle missing value method, feature selection technique and classification scheme for a particular clinical datasets.

For future work, the implemented algorithms for feature selection will be used as a predictor in the prognosis model for decision support system. The most important path is to design and create the appropriate predictive model and it is essential to think of what features should be selected and their precision and accuracy. Therefore, the challenge of uncertainty clinical data issue will arise to handle by probabilistic approach [2].

REFERENCES

[1] A. K. Tanwani, M. J. Afridi, M. Z. Shafiq, M. Farooq, " Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets," EvoBIO 2009, 128-139.

[2] P. Szolovists, "Uncertainty and Decisions in Medical Informatics," Methods of Information in Medicine, 34:111–21, 1995.

[3] M. S. B. Sehgal., I. Gondal and L. S. Dooley, "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data" *Bioinformatics*. Vol. 21 nop. 10, pp. 2417-2423, 2005

[4] N. Zhou, L. Wang, "A Modified *T*-test Feature Selection Method and Its Application on the HapMap Genotype Data," Genomics, Proteomics & Bioinformatics, 5(3–4), pp. 242–249, 2007.

[5] U. Fayyad, K. Irani, "Multi-interval discretization of continuous-valued attributes for classication learning," In: 13th International Joint Conference on Artificial Intelligencepp. 1022–1029, 1993.

[6] H. Liu, J. Li , L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," Genome Informatics, 13, 2002, pp. 51–60.

[7] L. Yu, H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," Machine Learning Research, 5, pp. 1205–1224, 2004.

[8] T. Jirapech-Umpai, S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," BMC Bioinformatics, 6, 148, 2005.

[9] W. D. Kim, H. K., Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids". Pattern Recognition Letters. vol. 25, pp. 1263-1271, 2004

[10] T. Kanungo, M. D Mount, S. N. Netanyahu, D. C. Piatko, R. Silverman and Y. A. Wu, "An Efficient *k*-Means Clustering Algorithm: Analaysis and implementation". IEEE Transactions and Pattern Analysis and Machine Intelligence. Vol. 24 (7), pp. 881-892, 2002

[11] K. Alsabti, S. Ranka, and V. Singh.,. "An Efficient K-Means Clustering Algorithm" pp.1-7 1997

[12] M.Omid, "Design of an expert system for sorting pistachio nuts through Decision Tree and Fuzzy Logic Classifier". Expert Systems with Appliaction. Vol. 38, pp. 4339-4347, 2011

[13] M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques". KDD workshop on text mining. pp.1-2, 2000

[14] F. Coetzee, "Correcting the Kullback-Leibler distance for feature selection", presented at Pattern Recognition Letters, 2005, pp.1675-1683.

[15] C.-N. Hsu, H.-J. Huang, D. Schuschel, "The ANNIGMA-wrapper approach to fast feature selection for Neural Nets," IEEE Transactions Systems, Man and Cybernetics, Part B, 2002, pp. 1–6.

[16] A.-N. Yahya, M. G. Kevin, Z. Jufen, G.F. C. John, L. C. Andrew, "Red cell distribution width: an inexpensive and powerful prognostic marker in heart failure," European Journal Heart Failure, vol. 11, pp. 1155–1162, 2009.

[17] Atherotech Diagnotics Lab, "Atherotech Panels," [Online], (URL http://www.atherotech.com/athdiagtests/atherotechpanels.asp), (Accessed 13 June 2011).

[18] D. W. Aha, R. L. Bankert, "A Comparative Evaluation of Sequential Feature Selection Algorithms," In: Fifth International Workshop on Artificial Intelligence and Statistics, pp. 1–7, 1995.

[19] S. Li, C. Liao, J. T. Kwok, "Gene Feature Extraction Using *t*-Test Statistics and Kernel Partial Least Squares," ICONIP, 3, pp. 11–20, 2006.

[20] L. Wang, F. Chu, W. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 40–53, 2007.[23]

[21] C. L. Bean, and C. Kambhampati., "Knowledge-Oriented Clustering for Decision Support". IEEE. pp. 3244-3249 2003.

[22] Z. Pawlak, "Rough Sets. International Journal of Computer and information Sciences". Vol. 11 (5), pp.341-356, 1982