International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 9, September - 2013

Design and Implementation Of An Efficient Relative Model in Cancer Disease Recognition.

Vikas Tiwari[1], Asst. Prof. T.D.Diwan[2], Asst. Prof Rohit Miri[3]
1 Dr. CVRU Bilaspur, 2 Dr. CVRU Bilaspur, 3 Dr. CVRU Bilaspur

Abstract – Cancer is a major public health problem in the India and many other parts of the world. Most of deaths in the India are due to cancer. In present time it is very difficult to identify relative patterns in cancer disease.

This paper presents an automatic and effective approach to identify hidden pattern of cancer disease is proposed using data mining technique based on association rules (AR) and clustering with fuzzy set theory and full attention are given on confidentiality and secrecy of patients.

This is being achieved through rigorous experimental design and in-depth quantitative studies. The expected outcome is the development of panels of biomarkers that will allow early detection of cancer and prediction of the probable response to therapy. Achieving these objectives requires high-quality specimens with well-matched controls, reagent resources, and an efficient process to confirm discoveries through independent validation studies

Index Terms — Fuzzy Set Theory, Clustering, association rule, point energy, cluster energy.

I. INTRODUCTION

Cancer is a class of diseases characterized by out-of-control cell growth. Cancer harms the body when damaged cells divide uncontrollably to form lumps or masses of tissue called tumors (except in the case of leukemia where cancer prohibits normal blood function by abnormal cell division in the blood stream). Tumors can grow and interfere with the digestive, nervous, and circulatory systems and they can release hormones that alter body function. Tumors that stay in one spot and demonstrate limited growth are generally considered to be benign.

There are over 100 different types of cancer, and each is classified by the type of cell that is initially affected.

More dangerous, or malignant, tumors form when two things occur:

1. a cancerous cell manages to move throughout the body using the blood or lymph systems, destroying healthy tissue in a process called invasion
2. that cell manages to divide and grow, making new blood vessels to feed itself in a process called angiogenesis.

Many things are known to increase the risk of cancer, including tobacco use, dietary factors, certain infections, exposure to radiation, lack of physical activity, obesity, and environmental pollutants.These factors can directly damage genes or combine with existing genetic faults within cells to cause cancerous mutations.Approximately 5–10% of cancers can be traced directly to inherited genetic defects.Many cancers could be prevented by not smoking, eating more vegetables, fruits and whole grains, eating less meat and refined carbohydrates, maintaining a healthy weight, exercising, minimizing sunlight exposure, and being vaccinated against some infectious diseases.

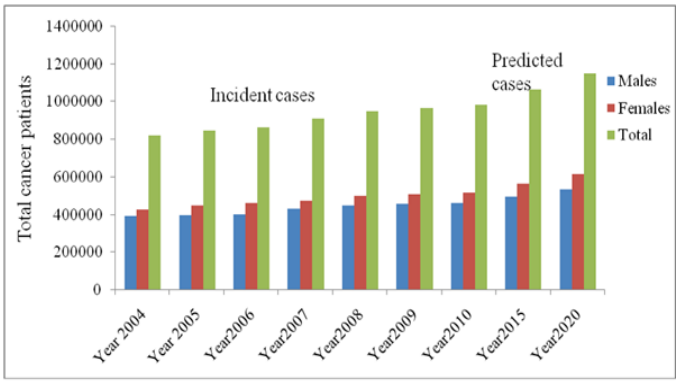

Figure 1: Year wise Incident cases and predicted Cases of cancer

Figure 1 represents year wise incident cases of cancer from year 2004 to till now and then predicated cases of cancer till now. It is clear from figure that cancer disease increasing rapidly in India in male and female both categories.

Cancer disease is classified in Environmental Cancer and Genetic cancer on the basis of it occurs due to family history or due to environmental factors.

Breast cancer and ovarian cancer may occur due to genetic mutation of BRCA1 and BRCA2 genes. It is identified that when a woman has genetic mutation of BRCA1 and BRCA2 genes she has 60 % life time risk of developing breast cancer. 5% of breast cancer patient are due to genetic mutation and 50 % risk of acquiring genetic mutation of BRCA1 and BRCA2 genes from mother. It is identified that when mother of a woman having breast cancer then developing breast cancer in that woman is 1.7 times higher than general population.

S NO	NAME OF CANCER	TYPE OF CANCER	SOURCES
1.	Breast Cancer	GENETIC	BRCA1,BRCA2 Genes
2.	Colon Cancer	GENETIC	APC Genes
3.	Ovarian Cancer	GENETIC	BRCA1,BRCA2 Genes
4.	Bladder Cancer	ENVIORMENTAL	Metal, Chlorination by-products , Aromatic amines , Ionizing radiation
5.	Bone Cancer	ENVIORMENTAL	Ionizing radiation
6.	Breast Cancer	ENVIORMENTAL	Ionizing radiation , Endocrine disruptors , Solvents , tobacco smoke
7.	Cervical Cancer	ENVIORMENTAL	Solvent
8.	Esophageal Cancer	ENVIORMENTAL	Metals
9.	Kidney Cancer	ENVIORMENTAL	Metals , Solvent
10	Laryngeal Cancer	ENVIORMENTAL	Metalworking fluids , Mineral oils , Natural fibers , Reactive chemicals
11	Liver and Biliary Cancer	ENVIORMENTAL	Metals , solvents , Reactive chemicals , Ionizing radiation , Polychlorinated biphenyls
12	Lung Cancer	ENVIORMENTAL	Metals , Solvents , Ionizing radiation , Reactive chemicals , Environmental tobacco smoke , Outdoor air pollution, Indoor air pollution , Natural fibers ,
13	Ovarian Cancer	ENVIORMENTAL	Pesticides , Ionizing radiation
14	Pancreatic Cancer	ENVIORMENTAL	Metals , Solvent , Reactive chemicals , Pesticides, Metalworking fluids , Mineral oils
15	Prostate Cancer	ENVIORMENTAL	Pesticides, Endocrine disruptors , Metallic dusts , Metalworking fluids, Combustion products, Aromatic amines, Metals , Pesticide,

Table1: Categorization of cancer disease

Table1 represents list of several cancer and type of that cancer and its sources

Colon cancer may occur due to mutation in the APC gene. Persons having genetic mutation of APC having 100 % risk of developing colon cancer in the late adulthood. It is identified that chance of acquiring genetic mutation of APC genes from parents is 50 %. There is a very small chance of developing cancer due to family history in lung cancer, prostate cancer and esophageal cancer. For genetic cancer it is difficult to identify that cancer occurs due to environmental factor or due to family history.

II. RELATED WORK

Attribute Based clustering for Feature Selection

Generally real world data have too many irrelevant features. To get better analysis of data and more accurate results irrelevant features are needed to remove from data set. Process of selecting attributes for further processing is too tedious task. In attribute based feature selection we use vertical fragmentation in data set. We divide the data set in to two clusters one cluster having relevant attributes and another cluster having irrelevant attributes.

Vertical Fragmentation

For vertical fragmentation we transpose the data set and group data set in to two clusters. Following clustering approach is used for fragmentation.

Process begins with transposing the data set and assigning two data points as a cluster centers. Then it measure distance of data point from all cluster center and assign data point to that cluster which having minimum distance. This means data points are added to that cluster for which it is closer to cluster center. Then for each cluster we compute means of all data points in that cluster this will be the new cluster center for that cluster. Again we reallocate data point on the basis of minimum distance from cluster center this process repeat until cluster energy and point energy remain unchanged.

Euclidean distance

We use Euclidean distance to measure distance between two data points which is defined as

$$\text{Dist}(i, j) = \| x_i - x_j \|$$

$$= \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \ldots (x_{ip}-x_{jp})^2}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

$j = (x_{j1}, x_{j2}, \ldots, x_{jp})$

Cluster Energy

Cluster energy is computed for each cluster. Cluster energy of any cluster is some of distance of all data points from cluster center in that cluster. Cluster Energy for any cluster is computed by following formula.

$$C_{i=\sum} \text{dis}(p_{i,}x_{ij})$$

Where i is cluster number and j is point number of cluster i.

Point Energy

Point energy is computed for each data points. Point energy of any data point is minimum distance of that data point from all cluster centers. We use following formula to compute Point energy of any data point.

$$P_i = \text{Min } (\text{dis } (P_{i,} m_j))$$

Where i represent point number m_j j^{th} Cluster center and j is cluster number.

Intuitionistic Fuzzy Set Theory

Fuzzy set is a set having element and its corresponding membership value. According to Fuzzy theory when an element X having membership value m(X) then its non-membership are computed by 1- m(X) but in Intuitionistic fuzzy set theory it is not true. When a person having smoking habit then supposes it membership value for caner prediction is suppose 0.6 then it is not necessary chance of not a cancer patient will be 0.4. We use medical expert knowledge to provide membership value and nonmember ship value of any attribute value of cancer patient.

α-cuts

Sometimes we need to have only those elements in Fuzzy set which member ship value is greater than some threshold $\alpha \in [0,1]$.Suppose Fuzzy set A having an element $x \in X$ and α threshold value is represented by α then for α-cut A_α of A is represented as below:

$$A_\alpha = \{x \in X, m(A(x)) >= \alpha\}$$

Fuzzy similarity based classifier used for feature selection are more effective compa re to principle component analysis approach for medical data. Fuzzy similarity measures are more effective to build classifier on medical data set [12].

III. METHODOLOGY

A. Data Collection

Data set is taken from SGPGI Lucknow. This data set contains total 13 attributes and 768 cancer patients' details. Records of some patients are incomplete and some attributes are irrelevant

B. Data Preprocessing

In data preprocessing too handle missing attribute value of data set. When any record of patient having more than two attributes missing we apply List wise deletion. In list wise deletion approach we simply omit those cases which having missing data from data set and analyze only remaining data .This approach often decreases the sample size available for analysis. When any record of patient having one or two missing value we apply Mean substitution approach. In this approach we replace the missing value by substituting a mean for the missing data, suppose we do not have Farthest Extension of tumor size, we replace the mean Farthest Extension of tumor size. By using mean substitution approach we make only a trivial change in the correlation coefficient and no change in the regression coefficient. After applying this approach to handle missing value 22 patient records are deleted from data set and 746 records left for further processing.

C. Attribute Based Feature Selection and Feature Extraction

Applying natural clustering for feature selection and feature extraction to build a k-nearest neighbor classifier on medical data gives more accurate result in comparison of principle component analysis approach. For feature selection and extraction group the heterogeneous feature in to group of homogeneous feature then extract features useful in classification technique [13].

Feature selection is needed to select only relevant features from data set to perform analysis and further processing for more accurate result of processing and better result of analysis. Suppose we have a data set having attribute set A (A_1,A2.....A_k) with k attributes and records of n patients (X_1,X_2.....X_n). Then we transpose data set and now data set having k records with n dimension.

Now we apply clustering on data set and group it in to two clusters. Initially we choose any two attributes as cluster center of the two clusters and compute point energy of all attributes assign each attributes to that cluster for which it has minimum distance. Then we compute Cluster energy of both cluster and divide the cluster energy by number of attribute that both clusters have this will be the new cluster center for both clusters this process will be repeated until cluster and point energy become unchanged. We propose the following algorithm for attribute based clustering for feature selection.

Attribute based clustering for Feature selection Algorithm

Step 1: Assign arbitrarily two attributes as cluster center for two clusters.

Step 2: Compute point energy of all attributes assign each attributes to that cluster for which it has minimum distance.

Step 3: Cluster energy of both clusters and divide the cluster energy by number of attribute that both clusters have this will be the new cluster center for both clusters.

Step 4: Repeat step 2 and step 3 until cluster energy and point energy remains unchanged.

After applying feature selection algorithm we get two clusters of attribute first cluster contains all relevant attribute and second cluster contains all irrelevant attributes. First cluster are selected for further processing.

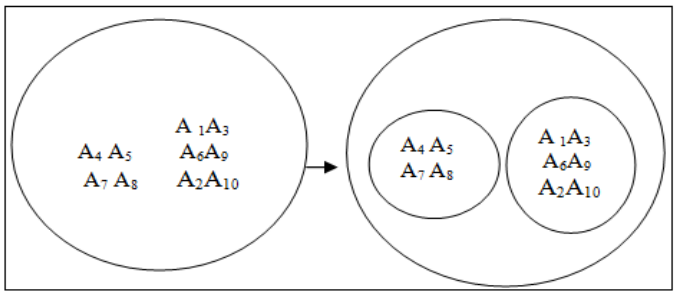

Figure 2: Clustering of attribute in two clusters

Figure two represents clustering of attributes in two clusters of relevant and irrelevant attribute clusters.

D. Categorization of data set

Cancer patient data set is categorized in three category of low risk patient medium risk patient and high risk patient using clustering approach as we apply for feature selection approach.

E. Rule mining

We need expert knowledge for fuzzification of the attribute value for each attribute for certain range of attribute value we determine membership value simultaneously nonmember ship value and modify the patient records by replacing the attribute value by their membership value and non membership value. Then we apply α-cut for α=0.6 Then we apply apriori algorithm for determination of association rules. Using this approach we get relative patterns between the attributes. These association rules can be used to built classifier and prediction of new patient class label weather they belongs from low risk, high risk or medium risk category.

Flow chart of overall approach is shown below

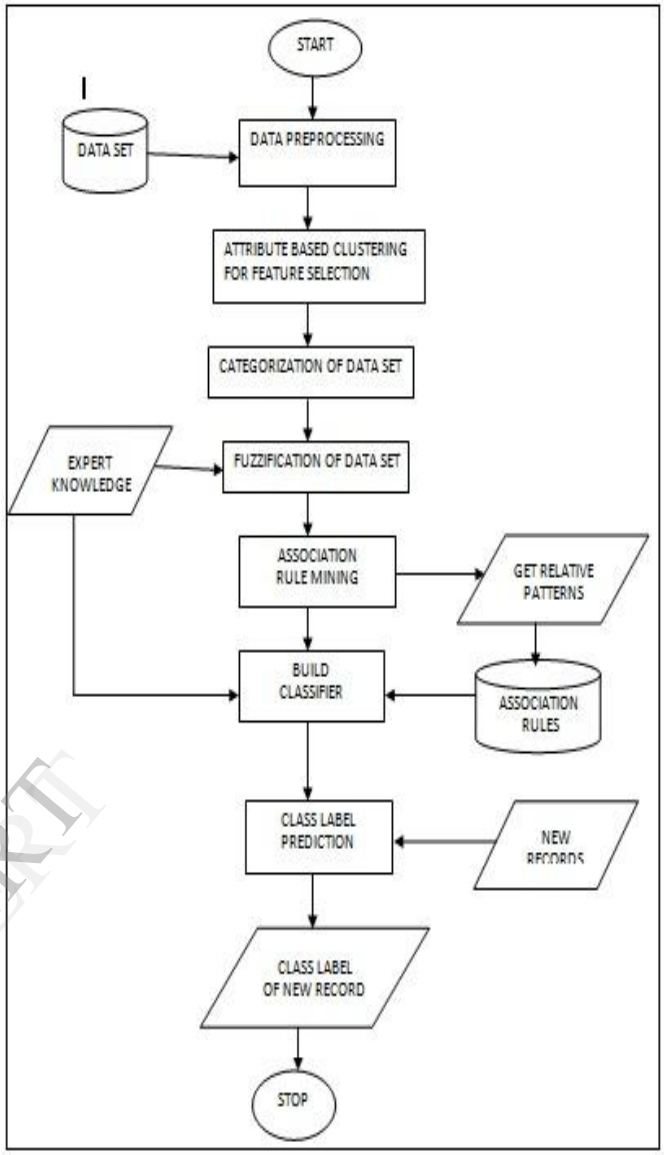

Figure 3: Proposed flow diagram

Figure 3 represents flow chart of our approach for determining relative patterns between attributes of cancer and predicting class label of patients.

IV. RESULT AND DISCUSSION

We apply attribute based clustering for feature selection. Data set contains total nine attributes out of nine seven attributes are selected as relevant attributes and two attributes are selected as irrelevant attributes.

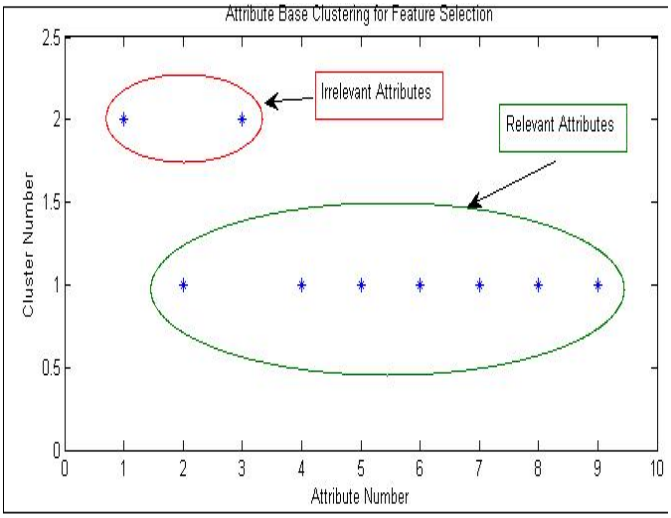

Figure 4: Result of Feature Selection

Figure 4 represents result of figure selection approach. When this result is compare with principle component analysis feature selection analysis approach we get similar result but our approach is more accurate to dealing with high dimension data set.

After feature selection we categorize data set in to three clusters of low risk medium risk and high risk patients

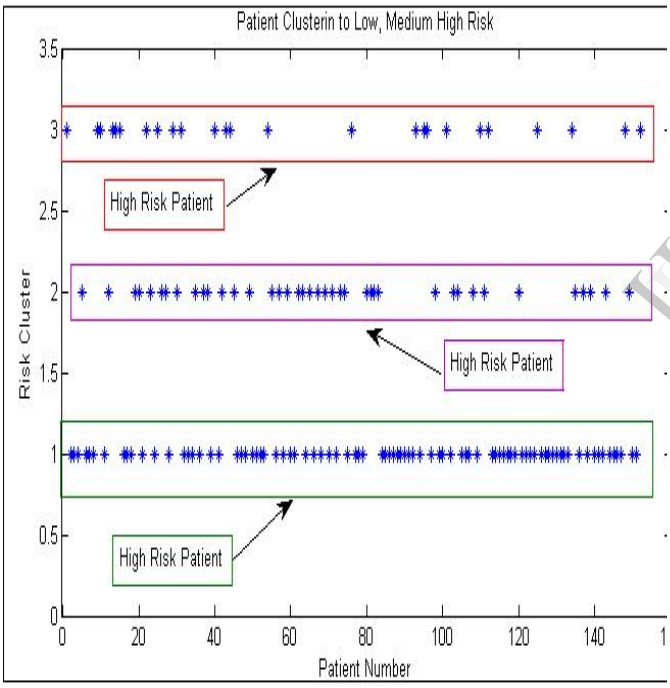

Figure 5: Result of Categorization of patient in to low medium and high Risk

Figure 5 represents result of categorization of patient in to low risk, high risk and medium risk cluster.

Relative patterns of cancer are derived from this approach. We explain our approach by example. Suppose cancer disease patients having five attributes with 5 patient records.

Patients Record

Tumor size	Lymph Node	Location	Place	Smoking habit
1.65	1	0	39	1
2.25	0	1	42	1
3.46	0	0	39	1
2.24	1	1	45	1
1.26	1	1	45	0

Table 2: Records of five patients

Table 2 represents a data set of five patients with 5five attribute dimension.

Step 1: We transformed the patient records using expert knowledge are given below.

Age		Lymph Node		Location		Place		Smoking habit	
S	US	S	US	S	US	S	US	S	US
0.4	0.7	0.6	0.5	0.6	0.3	0.6	0.4	0.6	0.4
0.5	0.5	0.6	0.5	0.4	0.5	0.6	0.4	0.6	0.5
0.6	0.3	0.6	0.3	0.6	0.3	0.6	0.4	0.6	0.5
0.4	0.7	0.6	0.3	0.4	0.5	0.5	0.6	0.6	0.4
0.2	0.8	0.5	0.5	0.4	0.5	0.7	0.4	0.4	0.5

Table 3: Patient record after applying Intuitionistic Fuzzy Set Theory

Table 3 represents data set after applying Intuitionistic Fuzzy Set Theory and using expert knowledge.

After applying α-cut for α=0.6 and representing useful value by 1 and remaining value by 0.

Age		Lymph Node		Location		Place		Smoking habit	
S	US	S	US	S	US	S	US	S	US
0	1	1	0	1	0	1	0	1	0
0	0	1	0	0	0	1	0	1	0
1	0	1	0	1	0	1	0	1	0
0	1	1	0	0	0	0	1	1	0
0	1	0	0	0	0	1	0	0	0

Table 4: Transformed Patient record for α-cut for α=0.6

Step 2: Now select only those attributes whose value is 1.then we get following transaction table.

Transactionid	Items
1	Age_2 $Location_1$ $place_1$ $Lymphnode_1$
2	$Place_1$ $smokinghabit_1$ $Lymphnode_1$
3	$Age_1$$Lymphnode_1$$location_1$ $place_1$ $smokinghabit_1$
4	Age_2 $Lymphnode_1$ $place_2$ $smokinghabit_1$
5	Age_2 $place_1$

Table 5: Transaction table

Step 3: Now we compute appearing time of each item as shown below.

Item	Appearing time
Age_1	1
Age_2	3
$Lymphnode_1$	4
$Lymphnode_2$	0
$Location_1$	2
$Location_2$	0
$Place_1$	4
$Place_2$	1
$smokinghabit_1$	4
$Smokinghabit_2$	0

Table 6 : Appearing time of each attribute

Step 4: Now for minimum support count 2 and minimum threshold 50 % and using apriori we get following following association.

Association Rule	Confidence
Smoking habit→Lumpphnode	100 %
Lumpphnode → Smoking habit	75 %
Place→ Smoking habit	66.6 %
Location→ Smoking habit	66.6%

Table 7 : Association Rules

IV. CONCLUSION

This analysis not only determine the hidden subjective result of cancer disease but also useful for hidden knowledge of other disease. It uses Fuzzy technique, point energy clustering, association rules mining and attribute based clustering for feature selection to determine association between the attribute of cancer. On the basis of these result we can predict the chances of happening cancer in % in any person. So it is very much fruitful for patients and doctors and experts of health to improve and maintain healthy environment.

V. REFERENCES

[1] Xiaobo Li1,2*, Sihua Peng3, Xiaosi Zhan1, Jinxiang Zhang1, Yueming Xu1, "Comparison of feature selection methods for multiclass cancer classification based on microarray data", 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI).

[2] Cheng-Fa Tsai†, Han-Chang Wu, and Chun-Wei Tsai, "A New Data Clustering Approach for Data Mining in Large Databases", 1087-4089/02 $17.00 © 2002 IEEE .

[3] Tapas Kanungo, David M. Moun, Nathan S. Netan, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[4] Razan Paul, Abu Sayed Md. Latiful Hoque, "Clustering Medical Data to Predict the Likelihood of Diseases", 978-1-4244-7571-1/10/$26.00 ©2010 IEEE.

[5] B. Liu, W. Hsu, and Y. Ma,"Integrating classification and association rule mining," SIGKDD, 1998.

[6] W. Li, J. Han, and J.Pei, "Cmar: Accurate and efficient classification based on multiple-class association rule", ICDM, 2001.

[7] T. Fadi, C. Peter, and Y. Peng, "Mcar: Multi-class classification based on association rule," IEEE International Conference on Computer Systems and Applications, 2005.

[8] X.Lin and J.Han,"Cpar: Classification based on predictive association rule," , SDM2003, 2003.

[9] G. Chen and H. et al,"A new approach to classification based on association rule mining," DECISION SUPPORT SYSTEMS, vol. 42, 2006.

[10] Raghvendra Mall, Prakhar Jain and Vikram Pudi, "PERFICT : Perturbed Frequent Itemset based Classification Technique.", 2010 22nd International Conference on Tools with Artificial Intelligence

[11] Rahul Isola, Rebeck Carvalho and Amiya Kumar Tripathy, "Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR and k-NN", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE - TITB-00346-2011.

[12] Pasi Luukka and Tapio Lepp¨alampi, "Similarity Classifier with generalized mean applied to medical data using different preprocessing methods", 0-7803-9158-6/05/$20.00 © 2005 IEEE.

[13] Mykola Pechenizkiy*, Alexey Tsymbal**, Seppo Puuronen*,"Local Dimensionality Reduction within Natural Clusters for Medical Data Analysis", Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05) 1063-7125/05 $20.00 © 2005 IEEE.

[14] Shoji Hirano Shusaku Tsumoto,"Strucutural Comparison and Cluster Analysis of Time-Series Medical Data".

[15] S.Cavuto E.Grossi,"The fuzzy nature of health and disease", 1 -4244-0363-4/06/$20.00 ©2006 IEEE.