A Novel Approach for Dimension Reduction in High Dimensional Medical Data

D. Ashok Kumar[*1], P.Velayudham[*2]
Department of Computer Science, Government Arts College, Trichy-620022,
Tamil Nadu, India

Abstract- Dimensionality Reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Dimensionality reduction techniques offer solutions that both significantly improve the computation time, and yield reasonably accurate clustering results in high dimensional data. Data mining technique called clustering is involved for medical data due to its high dimensionality and data sparsity. We proposed a Modern approach to capture and render vast quantities of clinical data during the health care process. Utilization of data analysis and data mining methods in medicine and health care is sparse. Medical data is one of the heavily used categorical type data. We proposed an algorithm can be very useful for reduce the possible symptoms of large medical and health care database to determining the best matching symptoms for the health disorders. Our objective is to search for clinically meaningful clusters of chronic illness, rather than clusters of patients. The results show that the clustering based symptoms observed approach is very efficient in terms of fast and timely decision making for medical representatives to help on time for the illness emergency patients.

1. INTRODUCTION

High dimensionality poses two challenges for unsupervised learning algorithms. First the presence of irrelevant and noisy features can mislead the clustering algorithm. Second, in high dimensions data may be sparse (the curse of dimensionality), making it difficult for an algorithm to find any structure in the data. To ameliorate these problems, two basic approaches to reducing the dimensionality have been investigated: feature subset selection (Agrawal et al., 1998; Dy & Brodley, 2000) and feature transformations, which project high dimensional data onto interesting subspaces (Fukunaga, 1990; Chakrabarti et al., 2002). For example, principle component analysis (PCA), chooses the projection that best preserves the variance of the data. The main drawback of PCA is that the size of the covariance matrix is proportional to the dimensionality of the data points. As a result, the computation of the eigenvectors might be infeasible for very high-dimensional data (under the assumption that $n > D$).

New Technology advances in medical science have led to revolutionary changes in medical research and the accumulation of a large volume of medical data that demands in-depth analysis. Business, financial, and scientific data can be easily modeled, transformed and applied formulas in contrast to medical data, whose underlying structure is poorly classified in mathematical terms. Hence, data mining techniques are required for processing the raw and sparse categorical medical data; clustering technique can be used to group the patients into different categories like normal, abnormal, intensively cared. Technological advancements in the form of computer-based patient records software and personal computer hardware are making the collection of access to health care data more manageable.

Clustering is used to group the patient illness using the profiles created with the close-ended questions. A common goal of the medical data mining is the detection of some kind of correlation. In this paper, we propose a new approach to solve this problem. Our approach utilizes the idea of dimension reduction. The right mining technique, for the right dataset is to select the best

Dimension Reduction technique. It can be classified into two categories: Feature extraction and Feature transformation. For textual data, the Feature extraction technique involves specifying the bounds on the term distribution in the collection, i.e., to consider only terms which occur between the predefined minimum and maximum number of documents in the collection. The Feature extraction methods are usually tailored according to the nature of the data, it is applicable in all data mining or process tasks. The Feature transformation techniques perform a transformation of the high dimensional vector data into a meaningful representation in lower dimension. It usually retains the most important dimensions (attributes), removes the noisy dimensions (irrelevant attributes) and reduces computational cost.

The commonly used dimensionality reduction methods include supervised approaches such as linear discriminant analysis (LDA) unsupervised ones such as principal component analysis (PCA) , and additional spectral and manifold learning methods. When the class label information is available, supervised approaches, such as LDA, are usually more effective than unsupervised ones such as PCA for classification. Linear discriminant analysis (LDA) is a classical statistical approach for supervised dimensionality reduction and classification. LDA computes an optimal transformation (projection) by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination. The optimal transformation in LDA can be readily computed by applying an eigen decomposition on the so-called scatter matrices. It has been used widely in many applications involving high-dimensional. However classical LDA requires the so-called total scatter matrix to be nonsingular. In many applications involving high-dimensional and low sample size data, the total scatter matrix can be singular since the data points are from a very high-dimensional space, and in general the sample size does not exceed this dimension. This is the well-known singularity or undersampled problem encountered in LDA.

Here we will review Dimensional Reduction in section II, discuss the proposed idea in section III, present some experimental results in section IV and conclude with section V.

2. DIMENSION REDUCTION

Data pre-processing is a crucial step of data mining that enables the abduction of irrelevant values from a dataset. An important aspect of data pre-processing is dimensionality reduction (DR). Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Dimensionality reduction aims at providing low-dimensional representations of high-dimensional data sets. Reduced dimensionality representations are useful for source coding, noise filtering, compression, clustering, and data mining. Specific examples include eigenfaces for face recognition, orthogonal decomposition in transform coding, and sparse PCA for cluster analysis. Two types of Dimensional reduction methods the Linear and Non-Linear that is subdivide into convex and non convex techniques. Convex techniques optimize an objective function that does not contain any local optima, whereas non-convex techniques optimize objective functions that do contain local optima.

The advantages of modern technology, the generation of multi-dimensional data has proceeded at an explosive rate in many disciplines. Data preprocessing procedures can greatly benefit the utilization and exploration of real data. There are a number of data preprocessing techniques including data cleaning, data integration, data transformation and data reduction. In a clinical data[1], average patients might have hundreds of different facts describing their current situation; in diabetes treatment about 200 facts are recorded. There is an urgent need to be able to aggregate this massive amount of information in a useful way. As a case study, a patient has four independent indicators describing the lifestyle, i.e., levels of smoking, exercise, alcohol, and weight. These could be combined into one aggregate measure indicating the overall lifestyle of the patient. In these four multidimensional terms reduced to just one. These data processing techniques can substantially improve the overall quality of the

patterns mined and/or the time required for the actual mining. Medical data[2] have the high volume due lot of factors caused in the disease. The medical representative identifies the disease with best matching factors with the help of dimensional reduction.

3. HEALTH CARE SYSTEM

3.1. EXISTING HEALTH CARE SYSTEM

New Technology advances in medical science have led to revolutionary changes in medical research and the accumulation of a large volume of medical data that demands in-depth analysis. Business, financial, and scientific data can be easily modeled, transformed and applied formulas in contrast to medical data, whose underlying structure is poorly classified in mathematical terms. Hence, data mining techniques are required for processing the raw and sparse categorical medical data; clustering technique can be used to group the patients into different categories like normal, abnormal, intensively cared. This is very hard to determine the patient disorder with related factors, for example a patient can be given one or more *diagnosis[3]*. These are only valid in specified time intervals, as the patient's condition changes over time. The set of possible diagnoses is given by a classification of diseases, e.g., the World Health Organization's ICD-10 standard. A diagnosis has an alphanumeric code, a descriptive text, and an associated period of validity. A specific diagnosis might be superseded by another as medical knowledge evolves, thus ending its validity, but for historical reasons it is important to keep it in the classification. Diagnoses are grouped into diagnosis groups, e.g., "Diabetes diseases" or "Pregnancy-related diseases," for overview purposes. One diagnosis can be a part of multiple groups, e.g., "Diabetes during pregnancy" can be a part of mentioned groups. The participation in the diagnosis groups of diagnoses can change over time, as the demands for grouping vary[9]. The groups also have an alphanumeric code and a descriptive text. A patient is treated according to a *protocol*, which is a formal description of how a treatment should progress. Different protocols are used depending on the characteristics, e.g., the age, of the patient. We will not go into the very complex internal structure of a protocol, but will just record a code, a text, and a period of validity. The protocol used for treating a patient may vary over time. An important indicator for the status of a diabetes patient is the condition of the *feet*, e.g., the blood circulation and presence of wound from time to time, the feet are photographed, and the pictures are stored along with the times they were taken. One of the most important measurements for diabetes patients is *HbA1c%*, which indicates the long-time blood sugar level and provides a good overall indicator of the patient's status during the recent months. This measurement is taken approximately every three months. For diabetes patients, a healthy lifestyle is even more important than normally, as it can literally make the difference between life and premature death. The life style factors are not measure regularly.

3.2. PROPOSED CLUSTERING BASED SYSTEM

We proposed an algorithm can be very useful for reduce the possible symptoms of large medical and health care database to determining the best matching symptoms for the health disorders[3]. Our objective is to search for clinically meaningful clusters of chronic illness, rather than clusters of patients. The results show that the clustering based symptoms observed approach is very efficient in terms of fast and timely decision making[1],[6] for medical representatives to help on time for the illness emergency patients.

The clustering based system to composed of diagnostic groups with the highest degree of association was the Metabolic Cluster[10] of the patients who had two or more chronic diseases, it fall in to one of the 15 possible subsets of the four diseases within this cluster: e.g., hypertension and hyper-lipidemia; hypertension and diabetes; hypertension, hyperlipidemia and diabetes; etc. The emergence of this cluster was not surprising, since it has been identified by Neurovascular, Dual Diagnosis[6] and mixed

Anxiety Depression. The diseases within this cluster have well-established epidemiological ties; diabetes (or underlying insulin resistance) is thought to be a risk factor for the other three diseases.

The cost to heath care systems for treating dual diagnosis patients is more than twice that for other psychiatric patients because to the increase the medical dimension[1],[2]. The K-means clustering algorithm remains one of the most popular and Standard clustering algorithms used in practice. The main reason for its usage is that it is simple to implement and it is fairly efficient. The k-means algorithm is given in the following section.

3.2.1. K-MEANS ALGORITHM

K-means is a partitioning method algorithm, which is detailed in Algorithm 1. This K-means algorithm has been discovered by several researchers across different disciplines, most notably. The popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally minimal solution. K-means algorithm works optimally with categorical and numeric data so that this is the best for clustering. It is simple and fairly fast results are easy to interpret and it can work under a variety of conditions hence it stand as the standard algorithm for clustering. K-means remains one of the most popular clustering algorithms used in practice. The fundamental idea is to find K average or mean values, about which the data can be clustered. The k-means algorithm is a simple iterative method to partition a given dataset into a user specified number of clusters K. K-means is initialized from some random or approximate solution for each iteration every object points to its nearest cluster and then objects belonging to the same cluster are averaged to get new cluster centroid. Each iteration successively improves cluster centroid until they become stable.

ALGORITHM 1: K-MEANS

The K-means algorithm is build upon the following operations:

Step 1: Choose initial cluster centroids $Z_1, Z_2, \ldots, Z_K$

randomly from the p points

$X_1, X_2, \ldots X_p$, $X_i \in R^q$

where q is the number of features/attributes

Step 2: Assign point X_i, i = 1, 2, …, p to cluster C_j,

j = 1,2,…,K

if and only if $\| X_i - Z_j \| < \| X_i - Z_t \|$,

t = 1, 2,…,K. and $j \neq t$.

Ties are resolved arbitrarily.

Step 3: Compute the new cluster centroids Z_1^*, Z_2^*,

…,Z_K^* as follows:

$Z_i^* = \frac{1}{l_j}\sum_{X_j \in C_j} X_i$ where i = 1, 2,…,K, and

l_j =Number of points in C_j.

Step 4: If $Z_i^* = Z_i$, i = 1, 2,…, K then terminate.

Otherwise $Z_i \leftarrow Z_i^*$ and go to step 2.

Except the first step, the other three steps are repeatedly performed in the algorithm until the algorithm converges. Note that in case the process does not terminate normally at Step 4, then it is executed for a maximum number of iterations.

The k-means algorithm converges when the assignments no longer change. The number of iterations required for convergence varies and may depend on p, this algorithm is linear based on the dataset size.

ALGORITHM 2: LDA+K-MEANS

The LDA+K-means algorithm is build upon the following operations:

Step 1: The same process up to the step-2.

Step 2: Compute the new cluster $Z_i^* = \frac{1}{l_j}\sum_{X_j \in C_j} X_i$ centroids $Z_1^*, Z_2^*, \ldots, Z_K^*$ as follows:

i).For each cluster j (1 <= j <= k), recalculate the centroids.

ii). Compute its distance from the centroid of the present nearest cluster.

iii).Divide the distance from sum of square error and take the average distance.

Step 3: If $Z_i^* = Z_i$, i = 1, 2,…, K then terminate. Otherwise $Z_i \leftarrow Z_i^*$ and go to step 2.

COMPUTATIONAL COMPLEXITY - Generally speaking, the algorithm is equivalent to t (LDA + K-means clustering), where t~=0 is the number of iterations of the algorithm to converge. Thus the computational complexity of the algorithm is O(pnt) for K-means clustering and O(p2nt) for LDA clustering where p is the dimension of data, n is number of data points. Finally, we note that when natural clusters in data are either close to spherical Gaussians or well separated, K-means clustering is a good model of the data distribution. LDA-Kmeans deals with the data distributions with dimensionality reduction which deviate from this situation.

DISTANCE MEASURES

Distance measure plays an important role in obtaining correct clusters. Distance measures assess similarity between the components of a vector, which in general can be formalized as

$$d_{ij} = \alpha \sum_{l=1}^{p} f(x_l, y_l) \quad (1)$$

where α is a coefficient depending on vectors x_l and y_l, and varying across similarity measures. The function f may represent the sum, difference, probability, or some other function applied to its arguments. Manhattan distance, Euclidean distance, Hamming are common functions. The selection of a distance measure plays an important role in partitioning the data set.

Depending on the type of attributes like numeric, continuous, categorical used similarity measures varies. Consider a m-dimensional Euclidean space, the distance between any two points, x = [x1, x2,…,xp] and y = [y1, y2,…,yp] is given as Common distance/ Squared Euclidean Distance in equation (1) i.e.L2 norm which is represented as Sq.Eucl.

$$\sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \quad (2)$$

and the Manhattan distance /Cityblock Distance is given in equation (3) i.e L1 norm which is the sum of absolute differences, represented as CB

$$\sum_{i=1}^{p} |x_i - y_i| \quad (3)$$

Brief introduction is obtained up to now about K-means algorithm & its performance measures. The new proposed idea to improve the clustering is done through Wiener Transformation which is explained in the following section.

WIENER TRANSFORMATION

A modern approach clustering technique based on Wiener Transformation is proposed here. Wiener Transform is efficient on large linear spaces. Usually Wiener Transformation is used in Image Restoration for noise removal filtering . In this paper, the data are transformed into real data using the Wiener Transformation, which is a statistical transformation. The approach is based on a stochastic framework. The transformed data is clustered using the K-means algorithm. Its main advantage is the short computational time it takes to find a solution so that the clustered data is very efficient compared to normal clustering.

The input for wiener transformation is stationary with known autocorrelation. It is a causal transformation. It is based upon linear estimation of statistics . The Wiener transformation is optimal in terms of the mean square error. The Wiener filter is a filter proposed by Norbert Wiener [5]. The syntax for Wiener filter is Y = wiener2 (X, [p q], noise) for two-dimensional image which is normally used for image restoration. Wiener2 function is used because input is a 2-dimensional matrix. This equation is used here for data mining task. The input X is a two-dimensional matrix and the output matrix Y is of the same size. Wiener2 uses a pixel wise adaptive Wiener method based on statistics estimated from a local neighborhood of each pixel. Wiener estimates the local mean μ in equation (5) and variance σ^2 in equation (6) around each pixel using the equations given below.

$$\mu = \frac{1}{pq}\sum_{n1,n2\epsilon\eta} X(n1,n2) \quad (4)$$

$$\sigma^2 = \frac{1}{pq}\sum_{n1,n2\epsilon\eta} X^2(n1,n2) - \mu) \quad (5)$$

where is the p-by-q local neighborhood of each pixel in the input matrix X. wiener2 then creates a pixelwise Wiener filter Y using these estimates are given in equation (7).

$$Y(n_1,n_2) = \mu + \frac{\sigma^2 - v^2}{\sigma^2}(X(n_1,n_2) - \mu) \quad (6)$$

where υ^2 is the average of all the local estimated variances. Here the data is pre-processed by transforming into real data using the Wiener Transformation for a vector. The transformed data is clustered using the above mentioned K-means algorithm.

EMPIRICAL STUDY AND RESULTS

Clustering is transformed to real domain by the linear Wiener transformation. Then K-means clustering algorithm is executed on the transformed data. The qualitative evaluation of quantitative results is done using the best metrics for clustering like Inter-cluster distance, Intra-cluster distance, Sensitivity and Specificity. Spect-heart disease dataset is taken from the UCI repository is considered here[7]. It is a good data set for testing ML algorithms; it has 267 instances that are described by 23 attributes. Data on cardiac Single Proton Emission Computed Tomography (SPECT) images[8]. Each patient classified into two categories: normal and abnormal. The results for Spect-heart disease dataset are tabulated in Table. The Inter-cluster distance μ is calculated for the Squared Euclidean Distance, City Block Distance and Hamming Distance for the Real dataset (AD) and the Wiener transformed dataset (TD).

Inter-cluster distance means the distances between different clusters, and it should be maximized i.e distance between their centroids.

The inter-cluster distance μ for K clusters C1,C2,…CK with centroids Zi, i=1...K is given in equation (8).

$$\mu(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{j=i+1}^{K} |Z_i - Z_j| \quad (8)$$

Table: 1 Inter-cluster Distance μ for SPECT-heart disease dataset

Distance Measure	μ_1		μ_2		μ_3		μ_4		μ_5		Average μ	
	RD	TD	RD	TD	RD	TD	RD	TD	RD	TD	RD	TD
Sq. Eucl	10.179	5.162	10.756	10.799	10.59	10.907	9.693	8.396	10.35	9.661	10.119	10.480
CB	8.579	11.272	9.642	10.333	9.139	9.635	9.939	9.806	8.838	11.316	9.227	10.472
HD	9.685	10.59	9.414	9.709	8.461	9.459	9.239	10.134	10.429	10.557	9.446	9.998

From Table 1, it is observed that the Inter-cluster distance μi i= 1,2,..,5 for various distance measures in TD outperforms the RD. Five executions are done here because K-means clustering varies based on the initialization of centroids. On an average Wiener transformed clusters are best compared to normal clusters. During K-means clustering the Squared Euclidean Distance, City Block Distance and Hamming Distance are used to prove the performance of clustering. Total average inter-distance for all types of distance measures of RD is 9.47656 and 10.22593 for TD that means 0.75 improvements for μ on an average for TD.

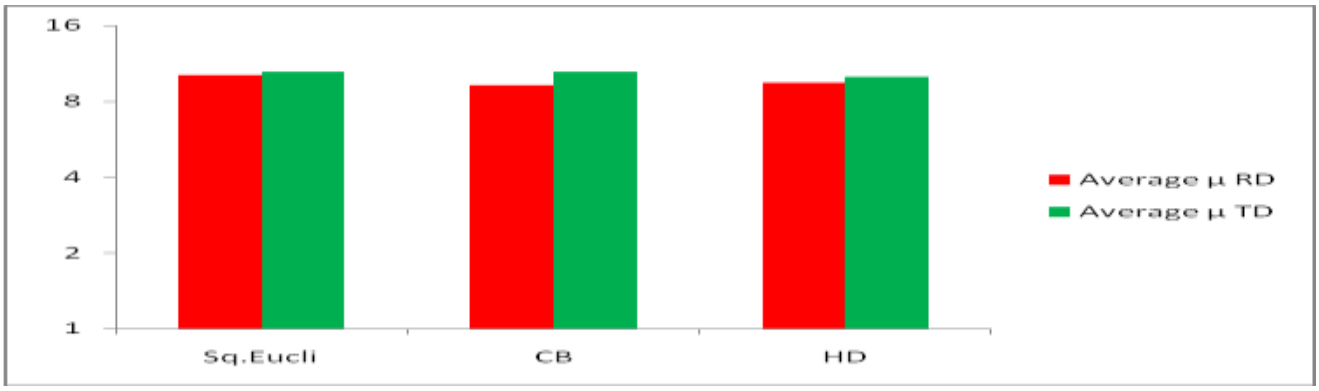

Fig. 1 Inter-distance Average performance analysis of various distance measures

Figure 1 depicts the overall average performance measures for K-means clustering for RD and TD with Squared Euclidean Distance, City Block Distance and Hamming Distance for clustering. City Block Distance has higher efficiency.

Intra-cluster distance is the sum of distances between objects in the same cluster, and it should be minimized.

The intra-cluster distance υ for K clusters C1, C2, .. CK centroid Zi, i=1..K is given in equation (9).

$$\upsilon(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{X_j \in C_i} |X_j - Z_i| \quad (9)$$

The Intra-cluster distance υ is calculated for the Squared Euclidean Distance, City Block Distance and Hamming Distance for the Actual dataset and the Wiener transformed dataset. The results are tabulated in Table 2 for Spect-heart disease dataset.

Table: 2 Intra-cluster Distance υ for Spect - heart disease dataset

Distance Measure	υ1		υ2		υ3		υ4		υ5		Average	
	AD	TD	AD	TD	AD	TD	AD	TD	AD	TD	AD	TD
Sq. Eucli	3.909	3.765	3.952	3.697	3.899	3.75	3.900	3.723	3.842	3.762	3.9006	3.732
CB	3.609	3.526	3.693	3.666	3.707	3.687	3.704	3.694	3.715	3.526	3.6862	3.620
HD	3.742	3.899	3.826	3.699	3.803	3.731	3.825	3.794	3.849	3.715	3.8407	3.736

The Table 2 shows that the Intra-distance υ for Squared-Euclidean distance in TD outperforms RD. Total average intra-distance for RD is 25.753 and 31.72437 for

TD that means 5.97 improvements for υ on a average. Wiener transformation clustering is best compared to clustering is depicted in the Fig 3.

Fig. 2 Intra distance average performance of various distance measures

Fig 2 depicts the average intra-cluster distance measures of Squared Euclidean Distance, City Block Distance and Hamming Distance. For Squared Euclidean Distance,and Hamming Distance TD has better performance. Considering CB, intra-distance is somewhat same.

4.1 Statistical Measures

Sensitivity and Specificity are statistical measures used in medical field; same measures are used here for evaluating wiener transformed data clustering. Sensitivity measures the ability of test to be positive when the condition is actually present, or how many of the positive test examples are recognized. A sensitivity of 100% means that the test recognizes all actual positives.

$$Sensitivity\ (SE) = \left(\frac{TP}{TP+FN}\right) * 100 \quad (10)$$

Specificity measures the ability of a test to be negative when the condition is actually present, or how many of the negative test examples are excluded. A specificity of 100% means that the test recognizes all actual negatives.

$$Specificity\ (SP) = \left(\frac{TN}{FP+TN}\right) * 100 \quad (11)$$

Predictive accuracy gives an overall evaluation.

$$Predictive\ Accuracy(PA) = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) * 100 \quad (5)$$

Where

TP - Number of True Positives

TN - Number of True Negatives

FP - Number of False Positives

FN - Number of False Negatives

Table: 3 Sensitivity, Specificity, Predictive Accuracy Value for Spect-heart disease dataset

Distance Measure	Sensitivity (%)		Specificity(%)		Predictive Accuracy Value (%)	
	RD	TD	RD	TD	RD	TD

Sq. Eucli	95	98	95	94	95	95
CB	93	93	98	97	94	95
HD	97	94	92	98	93	94

Sensitivity and Specificity for TD are always higher than the RD which is observed from Table 3. It means Wiener Transformation clustering is very effective.

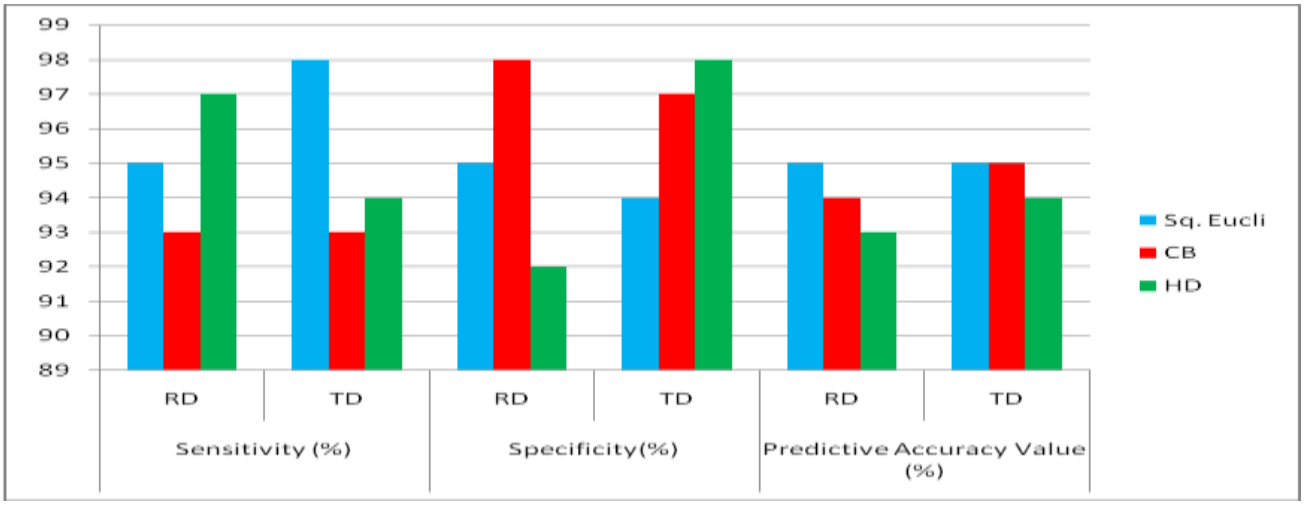

Fig 3. Average Performance of Sensitivity and Specificity of various distance measures

Fig 3 depicts the average performance of Sensitivity, Specificity and Predictive Accuracy Value for the Real data and the transformed data clustering with Squared Euclidean Distance, City Block Distance and Hamming Distance measures.

5. CONCLUSION

A novel approach has been proposed for decision making in medical treatment for emerging patient. Here dichotomous variables are grouped by clustering. The complex data is clustered by transforming the data into real cluster using the linear Wiener Transformation. The transformed data is clustered using the K-means algorithm, since k-means algorithm works fine in real domain it will give better results. The proposed technique empirically works well for finding good clusters. It is also observed that the statistical measures such as sensitivity and specificity is optimized and positive predictive is better in its average performances. Experiments are performed and the results show that clustering data using the Wiener transformation is more efficient than normal clustering with K-means.

REFERENCES

[1] Robert Leaman, Zhiyong Lu, Automated Disease Normalization with Low Rank Approximations. Processing (BioNLP 2014), pages 24–28,Baltimore, Maryland USA, June 26-27 2014.

[2] Laure Martin, Delphine Battistelli, Thierry Charnois, Symptom recognition issue, Processing (BioNLP 2014), pages 24–28,Baltimore, Maryland USA, June 26-27 2014.

[3] Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform, 47, 1-10.

[4] Doğan, R. I., & Lu, Z. (2012a). An improved corpus of disease mentions in PubMed citations. In Proceedings of the ACL 2012 Workshop on BioNLP, 91-99.

[5] Doğan, R. I., & Lu, Z. (2012b). An Inference Method for Disease Name Normalization. In Proceedings of the AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text, 8-13.

[6] Kavishwar B. Wagholikar, Manabu Torii, Siddartha R. Jonnalagadda and Hongfang Liu. 2013. Pooling annotated corpora for clinical concept extraction. Journal of Biomedical Semantics, 4:3.

[7] Uniqueness of Medical Data Mining (2002) by Krzysztof J. Cios , William Moore.

[8] Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M. & Goodenday, L.S. "Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis" Artificial Intelligence in Medicine, vol. 23:2, pp 149-169,Oct,2001.

[9] Centers for Medicare & Medicaid Services. 2014. General Equivalence Mappings. Documentation for Technical Users. Electronically published at cms.gov.

[10] Kavishwar B. Wagholikar, Manabu Torii, Siddartha R. Jonnalagadda and Hongfang Liu.. Pooling annotated corpora for clinical concept extraction. Journal of Biomedical Semantics, vol 4:3 Nov-2013.