Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SCALE TO DISEASE PRONENESS (SDP) & SCALE TO DISEASE INEPTNESS (SDI): DESIGN OF HEURISTIC METRICS TO ASSESS HEALTH CONDITION TOWARDS HEART DISEASE PRONENESS

RAMANA NAGAVELLI[1] DR.C.V.GURU RAO[2]
[1]Department of Computer Science and Engineering, Kakatiya University, Warangal, India
[2]Professor, SR Engineering College, JNTUH, Warangal, India
Email: [1]ramanauce.ku@kakatiya.ac.in [2]guru_cv_rao@hotmail.com

ABSTRACT:

The early stage diagnostics of heart disease is a challenging task. The interdependent and high complexity characteristics and factors related to heart disease combined with human constraints contribute towards the necessity of intelligent medical systems. In this paper we design heuristic metrics for predicting patterns of heart disease, the Scale to Disease Proneness (SDP) metric and Scale to Disease Ineptness (SDI) metric. The prediction system to be efficient in the performance and the scalability requirements has to select an optimal set of attributes from the data for which in our approach we make use of canonical correlation analysis. The test outcomes shown that the heuristic scales SDP and SDI devised here in this paper delivered optimal performance towards predication accuracy, also scalable and robust in the context of computational and process complexity. The approach tends to deploy easily to focus according to individual risk levels of the disease.

Keywords: *Health Mining, DDP, SDP, SDI, Disease prediction, Decision Support System, Machine Learning*

1. INTRODUCTION

The disease diagnostic decisions dependence on medical Decision Support Systems (DSS) [11] [3] has been augmented by the development of complicated disease-related insights dependent on numerous minute details, fast decision making requirements necessary for most critical cases, meeting the efficient care as well as enhanced value requirements of healthcare consumers, constraints of medical professionals having sufficient time, knowledge to handle each individual cases, the basic human limitations of applying the vast amount of data that may be applicable to any case, and the cost, time factors associated with training every medical professional.

The patterns and relationships hidden in huge databases [8] of mobile, internet, and medical data, extracted with data mining [6] techniques serve different industry information [7] requirements. The data mining approach combines techniques of database, artificial intelligence (AI), machine learning (ML), statistics and other tools [5]. The database is stored, organized, represented, and retrieved with disease diagnosis systems. The huge unordered data is categorized and classified into specific classes with algorithmic techniques of AI, NN, ML etc. The data mining strategies are modeled for different requirements such as, clustering, classification, pattern discovery, building predictive models, etc. however classification and prediction are the most primary functionalities widely applied. In classification feature selection techniques based on supervised and unsupervised learning model the data into groups for better analysis and comprehension. The prediction approach based on patient's available data is able to forecast the associated disease risk factors. The

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

strategy is implied to support decisions in all stages of diagnosis, prognosis and therapy processes.

The data mining strategies are applied with two techniques; supervised learning which uses machine learning algorithms to learn, train and classify the unorganized data with a training set, and unsupervised learning where no training set is used by the algorithms. The data mining techniques such as, bagging algorithm, naïve bayes (NB), and decision tree (DT), neural network (NN), kernel density (KNN), and support vector machine (SVM) both individually and in combination are used in the diagnosis and prediction process. The strategies are implemented with medical systems offering diagnostic support, for large volume of disease data, to assess the complex interdependence between various disease related factors, cause effect symptoms, and the true factors leading to the disease. The models of classification are able to classify the data without any hierarchy into specific labels, and the models of prediction show prediction functions of continuous-value [9]. The different data mining methodologies have varying accuracy levels in, processing the data, retrieving relevant information and predicting the disease.

The heart disease [4] diagnostics process is effective and simpler based on an analysis using statistics and data mining approaches. The attributes of heart disease have complex relationships and involve unique challenges in understanding if a patient's health characteristics imply causality. The modelling of heart disease-related data helps in generating parameters of risk factors enabling early stage diagnosis of high risk patients and timely administration of effective treatment. Several such analyses have studied the reasons of heart diseases and found that heart disease is mostly caused due to, heart attack, stroke or chest pain. The heart related problems which can be treated with the data mining strategies and techniques are, coronary heart disease (CHD) or heart attack [12], cerebrovascular disease (CVD) or stroke, hypertensive heart disease due to elevated blood pressure (hypertension), peripheral artery disease (PAD) due to reduced blood flow because of narrowing of arteries, rheumatic heart disease (RHD) or valve damage due to infections, coronary heart disease (CHD) or the heart defects that occurs during birth, and heart failure due to the heart muscles unable to pump enough blood to the heart, Inflammatory heart disease due to inflammation of the heart muscles. The database of heart disease obtained from the medical reports involves various elements having high number of attributes and our study for dimensionality reduction, attributes optimization and diagnostic process simplification, proposed the diseased heuristic scales SDP and SDI. The heuristic scale approach finds applications in financial forecasting, credit analysis, etc. and our study from the analysis of huge medical record database finds heuristic scales, with optimal attributes and searches for hidden knowledge among the relationships and explores the inadequately evaluated relationships effects.

2. RELATED WORK

A recent study states in the year 2010, considering all the reasons of deaths [2] in the world, the contribution of cardiovascular diseases (CVDs) was greater comparatively. The study also finds by the year 2030 about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [1] of which cardiovascular related diseases that are on the rise due to, alcohol, consumption, use of tobacco, lack of exercise, diet, etc. [3] would account for third of all the deaths [2]. Medical and health care professionals capability of accurately diagnosing heart disease is multiplied with the implementation of strategies, and techniques of data mining such as, classification, clustering, regression, artificial intelligence (AI), neural networks (NN), association rules (AR), decision trees (DT), genetic algorithm (GA), support vector machine (SVM), K-nearest neighbor (KNN) etc.

A study comparing the performance of data mining techniques, NB Classification, DT, Artificial Neural Networks (ANN) in the training, prediction of heart diseases finds different ways of improving the prediction performance. The study gathers and analyses patient records related to heart diseases, with 5 class labels and 15 attributes. The process uses, K-means clustering algorithm [13] for pre-processing and clustering, and Association Rules or ARM in the prediction process. The data mining, analysis and prediction tests results show, ANN based approach Artificial

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Neural Network based on Multi-layer Perceptron (MLP) with back propagation learning algorithm has more efficient performance of disease prediction in comparison to the other algorithms considered. The other approaches tested, K-Nearest Neighbor, a basic Neural Networks approach, clustering based Naïve Bayes classification approach, show lower predictive capabilities [14].

However a reduction of the test data and number of class labels shows better performance using, Decision Tree (DT) and a comparable precision sometimes is obtained with Bayesian classification. In this test, 909 heart diseases patient records are considered where the dataset is divided into two sets equally, training dataset comprising of 455 records and testing dataset comprising of 454 records. In the learning process 13 regular attributes are used [12] with 2 class labels "Heart Disease" and "No Heart Disease". The prediction process using the algorithms DT and NBC based on Weighted Associative Classifier (WAC) attains accuracy of 81.51% maximum. The accuracy is improved to 99.2% with DT technique, compared to other techniques when the data size is reduced and the attributes reduced to 6 from 13 with genetic algorithm (GA).

A prediction system for predicting heart attack based on algorithm ID 3, by Hnin Wint Khaing [15] represents various risk levels as a decision tree. In the preprocessing stage, K-means clustering algorithm is used to cluster the heart disease database and mine associated data of heart attack. The further mining of the data with MAFIA (Maximal Frequent Item set Algorithm) yields recurring patterns of heart attack. The algorithm ID3 used as training algorithm depicts with a decision tree the heart attack associated risk levels. The important patterns selected are used to train the ML algorithm for improving the disease prediction efficiency. The tests performed on the system for predicting heart attack show the approach is effective in accurately predicting the risks of occurrence.

A statistical technique Canonical Correlation Analysis (CCA) [21], [22] of earlier times has become widely used in the previous 10 years due to its reasonably worthy data analysis application results providing better understandable insights into the problems. The approach measures the linear relationships existing amongst any 2, X and Y multidimensional datasets with auto-covariance's and cross-covariance's matrices of second-order.

The prediction of the risk levels of heart attack has been studied by Chaurasia and Pal in the paper [16]. The study is based on the database obtained from UCI machine learning heart disease repository which comprises of 4 data sets one each from, Cleveland Clinic Foundation of United States, Hungarian Institute of Cardiology, V.A. Medical Center, California and University Hospital, Switzerland. The database is tested with data mining techniques such as, Naïve Bayes, J48 decision tree and Bagging based on 11 significance attributes. The test results demonstrate, bagging based approach compared to techniques of Naive Bayes and J48, has more precise performance and effective prediction of heart attack.

The approaches previously devised by us, Degree of Disease Possibility (DDP) [17] and Degree of Disease Possibility under Feature Correlation (DDP-FC) [18], are strategies of statistical assessment which differ from the contemporary models and are devised primarily for predictive accuracy and scalable requirements. From the development of these previous models by us we design new metrics and test with the database CHDD (Cleveland Heart Disease Dataset). In this paper we introduce the heuristic metrics, Scale to Disease Proneness (SDP) and Scale to Disease Ineptness (SDI).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. DEFINING A HEURISTIC SCALE TO DISEASE PRONENESS (SDP) AND SCALE TO DISEASE INEPTNESS (SDI)

3.1 Dataset Preprocessing

The heart disease diagnosed patient record contains 76attributes with values of type continues and categorical. The given dataset [19] is formed with patient records with 14 attributes that are considerable and benchmarked in earlier research [15,16, 17, 18]. Along the side of the heart disease patient records considered for experiments, we also opted to the patient records that are noticed as normal and not affected by heart disease. In regard to facilitate the optimization process devised here, the values of the attributes in the given dataset should be numeric and categorical. Henceforth, initially we convert all alphanumeric values to numeric values and then the continuous values to be converted to categorical.

Table 1: Description Of Dataset Attributes [20]:

Attribute ID	Attribute of Complete Record	Description	Value state of the Attrib
1	#3	age: patient age	(in years)
2	#4	sex: gender	(1 = male; 0 = female)
3	#9	cp: chest pain type	(Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4	#10	trestbps: resting blood pressure	(in mm Hg on admission to the hospital)
5	#12	chol: serum cholestoral	(in mg/dl)
6	#16	fbs: fasting blood sugar > 120 mg/dl	(1 = true; 0 = false)
7	#19	restecg: resting electrocardiographic results	(Value 0: normal, Value 1: having ST-T wave abnormality [T wave inversions and/or ST elevation or depression of > 0.05 mV], Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
8	#32	thalach: maximum heart rate achieved	
9	#38	exang: exercise induced angina	(1 = yes; 0 = no)
10	#40	oldpeak: ST depression induced by exercise relative to rest	
11	#41	slope: the slope of the peak exercise ST segment	(Value 1: upsloping, Value 2: flat, Value 3: down sloping)
12	#44	ca: number of major vessels (0-3) colored by flourosopy	(0, 1, 2, 3)
13	#51	Thal	(3 = normal; 6 = fixed defect; 7 = reversable defect)
14	#58	num: diagnosis of heart disease (angiographic disease status)	(Value 0: < 50% diameter narrowing, Value 1: > 50% diameter narrowing)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A. The procedure to represent the alphanumeric values as numeric values and continuous values as categorical values

- Let consider each attribute with alphanumeric values, then list all possible unique values and list them with an incremental index that begins at 1.
- Replace the values with their appropriate index.
- Let consider each attribute with continuous values, and then partition them into set of ranges with min and max values, such that the records distributed evenly through all these ranges.

3.2 Attribute Optimization for Defining Scale to Disease Proneness

Let partition the preprocessed set of patient records based on their labels, such that the records labeled as normal is one set, records labeled as diseased is other set. Consider the unique values of each attribute values set $f_i v(NRS)$ in the resultant records-set (NRS) of records labeled as normal and their coverage percentage as $f_i v = \{f_i(v_1,c_1), f_i(v_2,c_2), f_i(v_3,c_3), f_i(v_4,c_4),....., f_i(v_j,c_j)\}$.

Further the attribute optimization for diseased patient records is done as follows:

- Let consider the records set $rs(NRS)$ contains records those labeled as normal.
- Let $f_i(DRS)$ be the attribute f_i of DRS and $f_i(DRS)_{vs}$ be the set of values assigned to that attribute in DRS
- Create an empty set $\overline{f_i(NRS)_{vs}}$ of size $|f_i(DRS)_{vs}|$, then fill it with values from $f_i v(NRS)$ according to their coverage percentage such that $|f_i(DRS)_{vs}| \cong |\overline{f_i(NRS)_{vs}}|$.
- This process is opted to prepare the attribute values vector $\overline{f_i(NRS)_{vs}}$ of each attribute f_i the NRS,
- This process should be applied for all attributes of the record-set and refer that resultant attributes with values as a set $\overline{NRS}$.
- The canonical correlation (see section 3.4) will be done further, which is between each attribute values set $f_i(DRS)_{vs}$ and $\overline{f_i(NRS)_{vs}}$ of DRS and $\overline{NRS}$ respectively.
- Further, the attributes of the DRS can be considered as optimal, which are having canonical correlation is less than given threshold or zero. Further we form a record set $ODRS$, which is having records with values of only attributes that are assessed as optimal through canonical correlation, and this record set $ODRS$ is used further to define the scale to disease proneness (SDP).

3.3 Attribute Optimization for Defining Scale to Disease Ineptness

As like as the process explored in section 3.2, consider the unique values of each attribute values set $f_i v(DRS)$ in the resultant records-set (DRS) with records labeled as normal and their coverage percentage as

$$f_i v(DRS) = \{f_i(v_1,c_1), f_i(v_2,c_2), f_i(v_3,c_3), f_i(v_4,c_4),....., f_i(v_j,c_j)\}.$$

Further the attribute optimization for diseased patient records is done as follows:

- Let consider the records set $rs(DRS)$ contains records those labeled as diseased.
- Let $f_i(NRS)$ be the attribute f_i of NRS and $f_i(NRS)_{vs}$ be the set of values assigned to that attribute in NRS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Create an empty set $\overline{f_i(DRS)_{vs}}$ of size $|f_i(NRS)_{vs}|$, then fill it with values from $f_i v(DRS)$ according to their coverage percentage such that $|f_i(NRS)_{vs}| \cong |\overline{f_i(DRS)_{vs}}|$.
- This process is opted to prepare the attribute values vector $\overline{f_i(DRS)_{vs}}$ of each attribute f_i the DRS,
- This process should be applied for all attributes of the record-set and refer that resultant attributes with values as a set $\overline{DRS}$.
- The canonical correlation analysis (see section 3.4) will be done further, which is between each attribute values set $f_i(NRS)_{vs}$ and $\overline{f_i(DRS)_{vs}}$ of NRS and $\overline{DRS}$ respectively.

Further, the attributes of the NRS can be considered as optimal, which are having canonical correlation value that is less than given threshold or zero. Further we form a record set $ONRS$, which is having records with values of only attributes that are assessed as optimal through canonical correlation, and this record set $ONRS$ is used further to define the scale to disease Ineptness (SDI).

3.4 Canonical Correlation Analysis

The multidimensional datasets X and Y are two data sets considered and the in-between linear relationships are established with the auto covariance's and cross-covariance's matrices of second-order with standard statistical technique CCA and the results offered are creditable and comprehendible results. The technique is based on finding 2 bases one each for the datasets X and Y, where the X and Y datasets matrix of cross-correlation becomes diagonal whereas, correlations of the diagonal is maximized.

The parameters for implementing the canonical correlations in CCA are studied in the paper [20], [21] where, X and Y data vectors should be of equal number; however the data vectors $x \in X$ and $y \in Y$ may have varying dimensions assuming the mean is zero. The canonical correlations computation is solved using the equations of eigenvector.

$$C_{xx}^{-1}C_{xy}C_{yy}^{-1}C_{yx}w_x = \rho^2 w_x$$

$$C_{yy}^{-1}C_{yx}C_{xx}^{-1}C_{xy}w_y = \rho^2 w_y \quad (1)$$

Here $C_{yx} = E\{yx^T\}$ where ρ^2 or Eigen values are square of canonical correlations and w_x and w_x or the Eigen vectors are normalized CCA basis vectors. The solutions to the equations which are considered are those equivalents to non-zero value whose number is equivalent to x and y vectors lesser dimensional value.

The method followed in various ICA and BSS techniques is also used here where, x and y data vectors if prewhitened the solution (1) could be simplified [22]. Following the process of prewhitening, the canonical correlations C_{xx} and C_{yy} are both converted to unit matrices. As $C_{yx} = C_{xy}^T$, Eq (1) is converted to,

$$C_{xy}C_{xy}^T w_x = \rho^2 w_x$$

$$C_{yx}C_{yx}^T w_y = \rho^2 w_y \quad (2)$$

As these equations are however really equations depicting the singular value decomposition (SVD) [24] of the cross-covariance matrix C_{xy}:

$$C_{xy} = U\Sigma V^T = \sum_{i=1}^{L} \rho_i u_i v_i^T \quad (3)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Here U and V represent orthogonal square matrices $(U^TU = I,\ V^TV = I)$ comprising of u_i and v_i representing singular vectors. In our approach, the singular vectors considered above are w_{xi} and w_{yi} representing basis vectors delivering canonical correlations. The matrices U and V, and the subsequent u_i and v_i singular vectors dimensionalities usually vary according to the varied dimensions x and y data vectors. The pseudo diagonal matrix

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \quad (4)$$

includes a diagonal matrix D comprising of singular values equal to non-zero and attached with zero matrices which makes the matrix Σ to be compatible with various dimensions of x and y. The non-zero singular values are basically the nonzero canonical correlations whose number is lesser than any of x and y data vectors dimensions if C_{xy} or the cross-covariance matrix has full rank.

3.5 Defining the Scale to Disease Proneness (SDP)

Let consider the medical records set $ODRS$ that formed due to canonical correlation analysis (see section 3.2).

Further, form a set $F(ODRS)$ such that

$$F(ODRS) = \{f_1(ODRS) = \{v_{11}, v_{12}, v_{13}, ...v_{1a}\},$$
$$f_2(ODRS) = \{v_{21}, v_{22}, v_{23}, ...v_{2b}\},$$
$$.$$
$$.$$
$$.$$
$$.$$
$$.,$$
$$f_i(ODRS) = \{v_{i1}, v_{i2}, v_{i3}, ...v_{ic}\}\}$$

Here in the above description $f_i(ODRS) = \{v_{i1}, v_{i2}, v_{i3}, ...\} \forall i = 1..n$ represents the optimal attribute f_i and the unique values $\{v_{i1}, v_{i2}, v_{i3}, ...\}$ assigned to that attribute of all the records in set $ODRS$.

Rank the each value v_{ij} of optimal attribute f_i, which is based on their coverage in the $f_i(ODRS)$.

Further represent each patient record $\{r_i \forall i = 1... \mid ODRS \mid \wedge r_i \in ODRS\}$ as a set $rs(r_i)$ with the respective rank of the value of each optimal attribute as follows:

$$r_i = \{f_1(v_j \forall j \in \{1.. \mid f_1 \mid\}), f_2(v_j \forall j \in \{1.. \mid f_2 \mid\}),$$
$$f_3(v_j \forall j \in \{1.. \mid f_3 \mid\}), ..., f_i(v_j \forall j \in \{1.. \mid f_i \mid\})\}$$

$$rs(r_i) = \{r(f_1(v_j \forall j \in \{1.. \mid f_1 \mid\})), r(f_2(v_j \forall j \in \{1.. \mid f_2 \mid\})),$$
$$r(f_3(v_j \forall j \in \{1.. \mid f_3 \mid\})), ..., r(f_i(v_j \forall j \in \{1.. \mid f_i \mid\}))\}$$

Here in this description r_i is a record that belongs to the $ODRS$, which is representing the set of respective values of the optimal attributes. The representation $f_i(v_j \forall j \in \{1.. \mid f_i \mid\})$ is the value v_j of optimal attribute f_i, and $\mid f_i \mid$ represents the size of all possible values to the attribute f_i. And the set $rs(r_i)$ is representing patient record r_i by the respective ranks of the values of the optimal attributes. The representation $r(f_i(v_j \forall j \in \{1.. \mid f_i \mid\}))$ is the rank of the value v_j of the attribute f_i.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Further, for each $rs(r_i)$, find the aggregate rank $ar(r_i)$ as follows, which is an average of ranks representing the respective values of the optimal attributes of the patient record r_i

$$ar(r_i) = \frac{\sum_{i=1}^{n} r(f_i(v_j \forall j \in \{1...|f_i|\})}{n}$$

The standards defined by ANOVA[24],

(i) The measure average reflects the centrality of the distribution, but not significant to consider it alone as representation of the distribution, since it is not considering the uniform distribution.
(ii) The standard deviation of these ranks represents the how they deviated from each other, which is also not confirming the distribution status.
(iii) The kurtosis [24] represents the state of uniform distribution. If kurtosis found to be platy-kurtic (kurtosis value less than three), then it is representing the uniform distribution.
(iv) Henceforth, the distribution with platy-kurtic value is significant to consider as uniform distribution.

Henceforth, we measure the kurtosis of each distribution and order them by their kurtosis from minimal to maximal. The kurtosis of the ranks of each patient record t_i is measured as follows:

$$\sigma_{ar(r_i)} = \sqrt{\frac{\sum_{i=1}^{n} (r(f_i(v_j \forall j \in \{1...|f_i|\}) - ar(r_i))^2}{n}}$$

$$m4 = \frac{\sum_{i=1}^{n} (r(f_i(v_j \forall j \in \{1...|f_i|\}) - ar(r_i))^4}{n}$$

$$g_{(r_i)} = \frac{m4}{\sigma_{ar(r_i)}}$$

Here in these equations $\sigma_{ar(r_i)}$ represents the variation observed between ranks of optimal attributes of a patient record r_t and $g_{(r_i)}$ represents the kurtosis observed between the ranks of those optimal attributes.

Further we consider the patient records with platy-kurtic distribution of the ranks, and then mean of the ranks of these records will be considered as a scale to assess the disease proneness.

$$\mu(ODRS) = \frac{\sum_{i=1}^{n} ar(r_i)}{n}$$

Here $\mu(ODRS)$ represents the mean of the aggregate ranks of n patient records of $ODRS$

$$SDP = \frac{\sum_{i=1}^{m} ar(r_i)}{m}$$

Here in the above equation SDP represents the scale to disease proneness, m represents the number of records with platy-kurtic rank distribution ($g_{(r_i)} < 3$) and having the rank greater than $\mu(ODRS)$.

The lower and upper bounds of the scale will be assessed as follows:

$$stdv_{ODRS} = \sqrt{\frac{\sum_{i=1}^{m} (ar(r_i) - SDP)^2}{m-1}}$$

Here in above equation the standard deviation of the aggregate ranks of all record in $ODRS$ is measured

$$SDP_{low} = SDP - stdv_{ODRS}$$

$$SDP_{upr} = SDP + stdv_{ODRS}$$

3.6 Scale to Disease Ineptness (SDI).

The scale that devised here in this section is aimed to assess the Ineptness state of the medical record. The Strategy that explored on medical

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

record set *ODRS* to define SDP (see section 3.5) is also applied on medical record set *ONRS* to devise Scale to Disease Ineptness (SDI). The process applied on *ONRS* is briefed here:

Form a set $F(ONRS)$ such that

$$F(ONRS) = \{f_1(ONRS) = \{v_{11}, v_{12}, v_{13}, ...v_{1a}\},$$
$$f_2(ONRS) = \{v_{21}, v_{22}, v_{23}, ...v_{2b}\},$$
$$.$$
$$.$$
$$.$$
$$.$$
$$.,$$
$$f_i(ONRS) = \{v_{i1}, v_{i2}, v_{i3}, ...v_{ic}\}\}$$

Here in the above description $f_i(ONRS) = \{v_{i1}, v_{i2}, v_{i3}, ...\} \forall i = 1..n$ represents the optimal attribute f_i and the unique values $\{v_{i1}, v_{i2}, v_{i3}, ...\}$ assigned to that attribute of all the records in set $ONRS$.

Rank the each value v_{ij} of optimal attribute f_i, which is based on their coverage in the $f_i(ONRS)$.

Further represent each patient record $\{r_i \forall i = 1... | ONRS | \wedge r_i \in ONRS\}$ as a set $rs(r_i)$ with the respective rank of the value of each optimal attribute as follows:

$$r_i = \{f_1(v_j \forall j \in \{1..|f_1|\}), f_2(v_j \forall j \in \{1..|f_2|\}), f_3(v_j \forall j \in \{1..|f_3|\}$$
$$, ..., f_i(v_j \forall j \in \{1..|f_i|\})\}$$

$$rs(r_i) = \{r(f_1(v_j \forall j \in \{1..|f_1|\})), r(f_2(v_j \forall j \in \{1..|f_2|\})), r(f_3(v_j \forall j \in \{1..|f_3|\}))$$
$$....., r(f_i(v_j \forall j \in \{1..|f_i|\}))\}$$

For each $rs(r_i)$, find the aggregate rank $ar(r_i)$ as follows, which is an average of ranks representing the respective values of the optimal attributes of the patient record r_i

$$ar(r_i) = \frac{\sum_{i=1}^{n} r(f_i(v_j \forall j \in \{1...|f_i|\})}{n}$$

According to the ANOVA [25] standards (explored in section 3.5), we measure the kurtosis of each distribution and order them by their kurtosis from minimal to maximal. The kurtosis of the ranks of each patient record t_i is measured as follows:

$$\sigma_{ar(r_i)} = \sqrt{\frac{\sum_{i=1}^{n} (r(f_i(v_j \forall j \in \{1...|f_i|\}) - ar(r_i))^2}{n}}$$

$$m4 = \frac{\sum_{i=1}^{n} (r(f_i(v_j \forall j \in \{1...|f_i|\}) - ar(r_i))^4}{n}$$

$$g_{(r_i)} = \frac{m4}{\sigma_{ar(r_i)}}$$

Here in these equations $\sigma_{ar(r_i)}$ represents the variation observed between ranks of optimal attributes of a patient record r_t and $g_{(r_i)}$ represents the kurtosis observed between the ranks of the optimal attributes of the patient record r_i

Further we consider the patient records with platy-kurtic distribution of the ranks, and then mean of the ranks of these records will be considered as a scale to assess the disease proneness.

$$SDI = \frac{\sum_{i=1}^{m} ar(r_i)}{m}$$

Here in the above equation SDI represents the scale to disease ineptness, m represents the number of records with platy-kurtic rank distribution ($g_{(r_i)} < 3$) and having the aggregate rank greater than $\mu(ODRS)$.

The lower and upper bounds of the scale will be assessed as follows:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\overline{\mu} = \frac{\sum_{i=1}^{m} \{ar(r_i) \exists g_{(r_i)} < 3\}}{m}$$

The above equation is finding the mean $\overline{\mu}$ of the aggregate rank of the records with platy-kurtic feature rank distribution.

$$stdv_{ONRS} = \sqrt{\frac{\sum_{i=1}^{m} \left(ar(r_i) - \overline{\mu} \right)^2}{m-1}}$$

Here in above equation the standard deviation of the aggregate ranks of all record in $ONRS$ is measured

$$SDI_{low} = SDI - stdv_{ONRS}$$

$$SDI_{upr} = SDI + stdv_{ONRS}$$

The Scale to Disease Proneness (SDP) and Scale to Disease Ineptness (SDI) that are assessed from the given medical records set for training will be used further to assess the scope of disease proneness or ineptness of a given medical record.

For a given medical record mr to be tested, as explored in section 3.5,

- Preprocess the given record and reform it as medical record mr_{SDP} with only values of optimal attributes of the SDP and similarly reform it as medical record mr_{SDI} with only values of optimal attributes of the SDI .
- To assess the Disease Proneness of the medical record mr_{SDP} , form $rs(mr_{SDP})$ with the respective rank of the value of each optimal attribute of SDP .
- To assess the Disease Ineptness of the medical record mr_{SDI} , form $rs(mr_{SDI})$ with the respective rank of the value of each optimal attribute of SDI .
- Find the aggregate rank of the mr_{SDP} as $ar(mr_{SDP})$ and also find the aggregate rank of the mr_{SDI} as $ar(mr_{SDI})$

Further the state of given medical record mr is assessed as follows:

$ar(mr_{sdp}) \le SDP \;\&\&\; ar(mr_{SDI}) \ge SDI_{upr}$	Ineptness
$ar(mr_{sdp}) \ge SDP \;\&\&\; ar(mr_{SDI}) \ge SDI_{upr}$	Proneness
$ar(mr_{sdp}) \le SDP \;\&\&\; ar(mr_{SDI}) \ge SDI$	Ineptness
$ar(mr_{sdp}) \ge SDP \;\&\&\; ar(mr_{SDI}) \ge SDI$	Proneness
$ar(mr_{sdp}) \le SDP \;\&\&\; ar(mr_{SDI}) \le SDI_{low}$	Proneness
$ar(mr_{sdp}) \ge SDP \;\&\&\; ar(mr_{SDI}) \le SDI_{low}$	Proneness
$ar(mr_{sdp}) \le SDP \;\&\&\; ar(mr_{SDI}) < SDI$	Proneness
$ar(mr_{sdp}) \ge SDP \;\&\&\; ar(mr_{SDI}) < SDI$	Proneness

Table 2: The Scale of Disease Proneness and Ineptness and assessing strategy

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Here in the table 2, all possible combinations of SDP and SDI and the impact of those combinations explored. Regardless of the mr_{SDI}, if mr_{SDP} is greater than SDP then the record confirmed to be Disease Prone. But in contrast, the medical record's disease ineptness is dependent of SDP, which is indicating that though the mr_{SDI} is more than the value of SDI, it's mr_{SDP} must be less than the SDP to conclude that the given medical record mr is scaled as disease ineptness. This may leads to slight increase in false positives in prediction but strictly avoids false negatives, which is an accuracy measurement of disease scope.

If $ar(mr_{sdi})$ is less than SDI and $ar(mr_{sdp})$ is greater than SDP, then record is said to be the effected by heart disease.

If disease impacts score of record mr to be tested is not in the range of disease impacts score of any attack, and hale record score is not in the range of hale record scope threshold boundaries, then the record said to be normal.

4. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The experiments were carried out on benchmarking dataset that explored in section 3.1. Initially partitioned the processed dataset into normal and diseased records and then the optimal attributes of diseased medical records and normal medical records were traced out, which is by using the process explored (see section 3.2, 3.3, and 3.4). Further the scale to Disease Proneness (SDP) and Scale to Disease Ineptness (SDI) were devised through the process explored (see section 3.5 and 3.6). The exploration of the input data and results were shown in table 3, 4 and 5. The visualization of the optimal scope of attributes for diseased and hale records can be found in fig1 and fig2.

Table 3: Statistics Of The Experiment Results

Total Number of Records	303
Range of fields count in a record	13
Records used for defining scale	80% (242 records)
Records used for performance analysis	20% (61 records)
Scale of Disease Proneness SDP observed	7.11324
SDP_{low} observed	5.575204
SDP_{upr} observed	8.651277
Scale of Disease Ineptness SDI observed	2.982372
SDI_{low} observed	2.35923
SDI_{upr} observed	3.605513

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Selected Attributes Of The Medical Records Labeled As Diseased Under Different Canonical Correlation Threshold

<0.04		<0.05		<0.051(0.06) (mean of the CC		<0.1		<0.2	
Attribute ID	CC value	ID	CC	ID	CC	ID	CC	ID	CC
1	0.000398	1	0.000398	1	0.000398	1	0.000398	1	0.000398
7	0.021013	4	0.044147	3	0.050546	2	0.107357	2	0.107357
8	0.029505	7	0.021013	4	0.044147	3	0.050546	3	0.050546
9	0.03851	8	0.029505	5	0.050081	4	0.044147	4	0.044147
12	0.021347	9	0.03851	7	0.021013	5	0.050081	5	0.050081
		11	0.042488	8	0.029505	7	0.021013	6	0.103026
		12	0.021347	9	0.03851	8	0.029505	7	0.021013
				11	0.042488	9	0.03851	8	0.029505
				12	0.021347	10	0.098274	9	0.03851
						11	0.042488	10	0.098274
						12	0.021347	11	0.042488
						13	0.099685	12	0.021347
								13	0.099685

<0.04		<0.06		<0.07585903 (mean of the cc vales) (0.08)		<0.1		<0.2 (yesits all)	
1	0.011762	1	0.011762	1	0.011762	1	0.011762	1	0.011762
5	0.004778	5	0.004778	2	0.069674	2	0.069674	2	0.069674
8	0.039421	8	0.039421	5	0.004778	3	0.082918	3	0.082918
11	0.021706	9	0.043843	8	0.039421	5	0.004778	4	0.190604
		11	0.021706	9	0.043843	8	0.039421	5	0.004778
		13	0.043684	11	0.021706	9	0.043843	6	0.177771
				12	0.066248	10	0.089583	7	0.144176
				13	0.043684	11	0.021706	8	0.039421
						12	0.066248	9	0.043843
						13	0.043684	10	0.089583
								11	0.021706
								12	0.066248
								13	0.043684

Table 4: Selected Attributes Of The Medical Records Labeled As Normal Under Different Canonical Correlation Thresholds

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

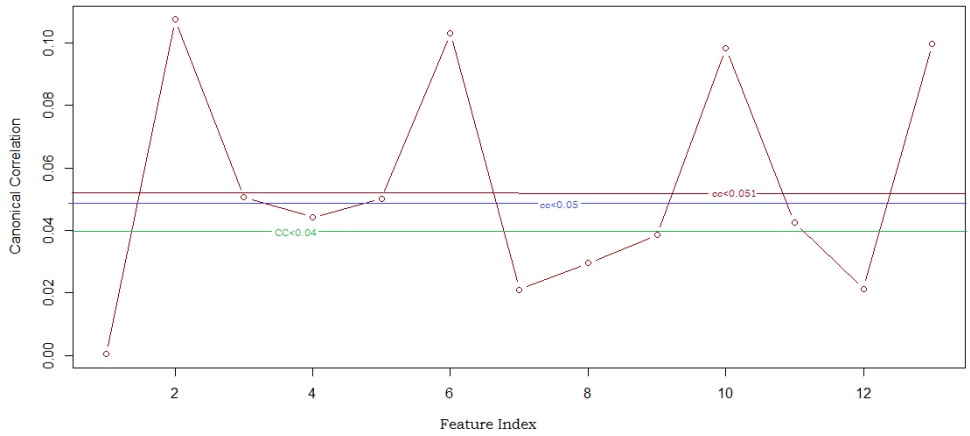

Fig 1: Attributes Of The Medical Records Labeled As Diseased And Their Optimality Under Divergent CC Thresholds

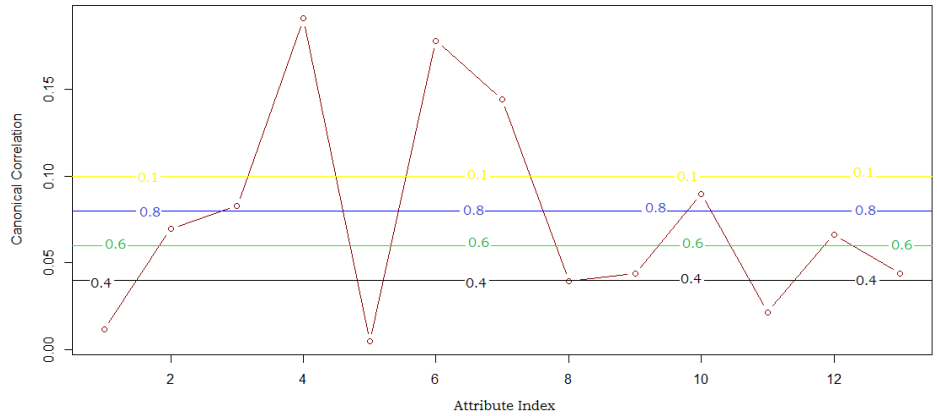

Fig 2: Attributes Of The Medical Records Labeled As Normal And Their Optimality Under Divergent CC Thresholds

4.1 Performance Analysis

The robustness and prediction accuracy of the scales *SDP* and *SDI* are assessed through 62 records, which are of the combination of 40 diseased and 22 normal records.

The prediction statistics are as follow:

The count of true positives are (records predicted as truly diseased) 40, the count of true negatives are (records predicted as truly normal) 20, the count of false positives are (records predicted as falsely diseased) 2 and the count of false negatives are (the records predicted as falsely normal) 0.

Since the results experimental results indicating that diseased record prediction is 40 out of given 40 records, hence prediction error towards diseased records are 0.0. The prediction accuracy of normal records is observed as 20 among the given 22 records, hence the prediction error ratio is approx. 0.09, which is to be negligible as in the track of medical practitioner, the record actually diseased should not be diagnosed as normal, in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

contrast a normal record can be suspected falsely as diseased and may recommend to further diagnosis strategies.

The prediction accuracy of the model devised here in this paper is explored through statistical assessment metrics called precision, recall and f-measure (see table 5). The value obtained for metric recall indicating that the devised model is highly robust and scalable towards assessing disease scope, and the precision also indicating that the prediction accuracy of the model is high and approximately it is 97%.

Precision	0.952381
recall	1
F-Measure	0.97561

Table 5: The Precision, Recall And The F-Measure Of The Predictions

The process time of the application is stable since the increase in number of optimal attributes is not influencing the process complexity (see fig 3)

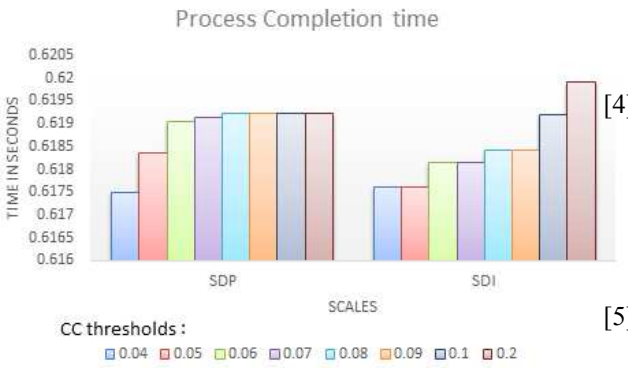

Fig 3: Process Completion Time Of SDP And SDI Under Divergent Optimal Attributes Selected Through Various Canonical Correlation Thresholds.

5. CONCLUSION:

This paper introduced a novel heuristic scale to assess the heart disease proneness of the given medical record of an individual patient. In regard to this, two heuristic metrics called Scale to Disease Proneness (SDP) and Scale to Disease Ineptness (SDI) is devised. In contrast to the existing benchmarking models, the proposed metrics are assessing the disease proneness and ineptness of the medical records SDP and SDI respectively. Further the combinations of these SDP and SDI values of the given medical record are used to assess the state of that medical record. The process opted to devise these metrics is initially finding the optimal attributes of the given diseased and hale medical records, which is done through the canonical correlation analysis. Further the medical records of diseased and hale with optimal attributes are used to assess the metrics SDP and SDI. The experimental results are optimistic and concluding the prediction accuracy and robustness. The future work can be the definition of fuzzy model to estimate the combination of SDP and SDI values.

REFERENCES:

[1] Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report, 2005.

[2] Global Burden of Disease. 2004 update (2008). World Health Organization.

[3] Srinivas, K.," Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[4] Yanwei Xing, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease", IEEE Transactions on Convergence Information Technology, pp(868 – 872), 21-23 Nov. 2007

[5] IBM, Data mining techniques, http://www.ibm.com/developerworks/opensource/library/ba-data-miningtechniques/index.html?ca=drs- , downloaded on 04 April 2013.

[6] Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/enus/virtuallabs/aa740409.aspx, 2007.

[7] Glymour C., D. Madigan, D. Pregidon and P.Smyth, "Statistical inference and data mining", Communication of the ACM, pp: 35-41, 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] Thurai singham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

[9] Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[10] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.

[11] C. Aflori, M. Craus, "Grid implementation of the Apriori algorithm Advances in Engineering Software", Vo lume 38, Issue 5, May 2007, pp. 295-300. A. J.T. Lee, Y.H. Liu, H.Mu Tsai, H.-Hui Lin, H-W. Wu, "Mining frequent patterns in image databases with 9DSPA representation",Journal of Systems and Software, Volume 82, Issue 4, April 2009, pp.603-618.

[12] Shanta kumar .Patil,Y.S.Kumaraswamy, "Predictive data mining for medical diagnosis of heart disease prediction" IJCSE Vol .17, 2011

[13] M. Anbarasi et. al. "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.

[14] Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011.

[15] V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.

[16] Quinlan J. Induction of decision trees. Mach Learn 1986; 1:81—106.

[17] Nagavelli, R.; Guru Rao, C.V., "Degree of Disease possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining," Recent Advances and Innovations in Engineering (ICRAIE), 2014 , vol., no., pp.1,6, 9-11 May 2014; doi: 10.1109/ICRAIE.2014.6909265

[18] RamanaNagavelli, Dr.C.V.Guru Rao; Degree of Disease Possibility by Feature Correlation (DDP-FC); International Journal of Advanced Computing, ISSN: 2051-0845, Vol.48, Issue.1; RECENT SCIENCE PUBLICATIONS

[19] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

[20] A. Rencher, Methods of Multivariate Analysis, 2nd ed., Wiley, 2002.

[21] M. Borga, "Canonical correlation: a tutorial", Linkoping University, Linkoping, Sweden, 2001, 12 pages. Available at http://www.imt.liu.se/~magnus/cca/tutorial/.

[22] A. Hyv¨arinen, J. Karhunen, and E. Oja, Independent Component Analysis. Wiley, 2001

[23] S. Haykin, Modern Filters. MacMillan, 1989.

[24] http://www.statsoft.com/textbook/anova-manova