COMPREHENSIVE FEATURE SELECTION BASED SUPPORT VECTOR MACHINE CLASSIFIER (CFS-SVM) FOR CLINICAL DATASET

S. KAVIPRIYA[1], DR.T. DEEPA[2]

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamil Nadu, India
Tamil Nadu, India
E-mail: kavipriya.vignesh@gmail.com[1], deepacs@srcw.ac.in[2]

ABSTRACT

Feature selection plays a significant role in any data mining research problem. In this research work, comprehensive feature selection is applied for selecting the attributes in the chosen PIMA Indian diabetes dataset. The comprehensive feature selection mechanism makes use of maximum significance pattern for selecting the most edifying features, which effectively distinguish between different classes of samples. By making use of the comprehensive feature selection mechanism, the degree of relevance between the feature variables and target variable is quantified, and the informativeness of features is determined according to the degree of the relevance score. In order to implement this mechanism, novel evidence metric, NEM, is employed to score the relevance degree of a candidate feature variable with respect to the target variable. Once after the feature selection is carried out, SVM classifier is applied for performing the prediction of heart disease among gestational diabetes patients in the dataset. Performance metrics such as sensitivity, specificity, true positive rate, false positive rate, precision, accuracy and time taken for feature selection are taken into account. The results are demonstrated with better performance.

Keywords: *Feature Selection, Data Mining, Gestational Diabetes, Accuracy, Time Taken, Feature Selection, Risk Prediction.*

1. INTRODUCTION

Data mining in health informatics is a computational process for mining facts from available clinical databases. Data mining is broadly classified as predictive data mining and descriptive data mining. Classification and prediction falls under the category of predictive data mining. Risk prediction of heart diseases is one among the thrust research area in the field of clinical decision support system. Knowledge based clinical decision support system (CDSS) usually consists of three components namely knowledge base, inference engine and mechanism to communicate. Non-knowledge based CDSS makes use of machine learning algorithms namely support vector machine, relevance vector machine, artificial neural network and extreme learning machine. These machine learning algorithms countenance computers to acquire from historical practices and/or find patterns in available clinical data. The non-knowledge based CDSS disregards the requirement for scripting rules and for proficient input. Nevertheless, these non-knowledge based CDSS based on machine learning possibly won’t elucidate the causes for their conclusions and hence known as “black boxes", for the reason that no expressive facts about the working will be distinguished by human scrutiny.

Almost all the clinical practitioners do not make use of such non-knowledge based CDSS directly for diagnoses due to the reliability and accountability reasons. On the other hand, they are used for design and development of post-diagnostic systems, for portentous patterns for clinicians to purview on it. In recent years, diverse data mining procedures have been used for decision support for heart disease risk prediction. Certain methods are portrayed in early which includes decision tree, neural network, and association rule mining, [R Alizadehsani et al.,2013, J Nahar et al.,2013, I.Kurt et al.,2008, M.G. Tsipouras et al.,2008]. Decision tree is easy to implement and interpret and it provides a tree-based classification for developing a predictive model according to independent variables [M.J. Berry and G. Linoff.,1997]. Decision tree seems to be one the precise algorithms among data mining tools.

.

2. LITERATURE REVIEW

Adaptive Genetic Fuzzy System (AGFS) was proposed by B. Dennis and S.

Muthukrishnan.,2014 for optimizing rules and membership functions for medical data classification process. To establish the efficiency of the proposed classifier the presentation of the anticipated genetic-fuzzy classifier was evaluated with quantitative, qualitative and comparative analysis. Hybrid Prediction Model (HPM) proposed by B.M. Patil et al.,2010 used Simple K-means clustering algorithm aimed at validating chosen class label of given and subsequently applying the classification algorithm to the result set. C4.5 algorithm was used to build the final classifier model by using the k-fold cross-validation method. J. Jabez Christopher et al.,2015 used a meta-heuristic approach called Wind-driven Swarm Optimization (WSO). The uniqueness of this work lies in the biological inspiration that underlies the algorithm. WSO uses Jval, a new metric, to evaluate the efficiency of a rule-based classifier. Rules were extracted from decision trees. WSO was used to obtain different permutations and combinations of rules whereby the optimal rule set that satisfies the requirement of the developer was used for predicting the test data.

J. Mattila et al.,2012 presented a novel generic clinical decision support system, which models a patient's disease state statistically from heterogeneous multiscale data, where the goal was to aid in diagnostic work by analyzing all available patient data and highlighting the relevant information to the clinician. Jamal Salahaldeen Majeed Alneamy and Rahma Abdulwahid Hameed Alnaish, 2014 intended to use the hybrid teaching learning based optimization (TLBO) algorithm and fuzzy wavelet neural network (FWNN) for heart disease diagnosis. The TLBO algorithm was applied to enhance performance of the FWNN. The hybrid TLBO algorithm with FWNN was used to classify the Cleveland heart disease dataset obtained from the University of California at Irvine (UCI) machine learning repository. Kalpana M and Kumar AS,2012 expressed the prominent features of the fuzzy expert system by applying the algorithm Fuzzy Assessment Methodology using K ratio, that was to diagnosis the diabetes Fuzzy Assessment Methodology using k ratio was developed. Fuzzy Expert System consists of following elements such as Fuzzification interface, Fuzzy Assessment Methodology using K ratio and Defuzzification interface.

Kamadi V.S.R.P. Varma et al.,2014 proposed a method to minimize the calculation of Gini indices by identifying false split points and used the Gaussian fuzzy function because the clinical data sets were not crisp, where the efficiency of the decision tree depends on many factors such as number of nodes and the length of the tree, pruning of decision tree plays a key role. Kindie Biredagn Nahato et al.,2016 proposed a classifier that combines the relative merits of fuzzy sets and extreme learning machine (FELM) for clinical datasets. The three major subsystems in the FELM framework were preprocessing subsystem, fuzzification subsystem and classification subsystem. Missing value imputation and outlier elimination were handled by the preprocessing subsystem.

Hybrid Intelligent System proposed by Manjeevan Seera et al.,2014 consists of Fuzzy Min–Max neural network, the Classification and Regression Tree, and the Random Forest model and its efficacy as a decision support tool for medical data classification was examined. Genetic Programming based method proposed by Muhammad Waqar Aslam et al., 2016 has been used for diabetes classification, which was also used to generate new features by making combinations of the existing diabetes features, without prior knowledge of the probability distribution. Medical Diagnosis System proposed by Vijaya K et al.,2010 predicted the severity of the cardiovascular diseases, where system was built by combining the relative advantages of fuzzy logic, neural network and genetic algorithm. The input variables that were non-discrete were fuzzified and fed as input to train the neural network.

3. PROPOSED WORK

3.1 Design of Algorithm

O. Kurşun et al.,2010 and C.O. Sakar et al.,2012 scrutinized the risk related to insignificant repetition in minimal-Redundancy-Maximal-Relevance (mRMR). To outline the issue, a basic graphical representation is given. In Fig. 1 there are three factors, including an objective variable v, a previously chosen feature a_1 and a candidate feature a_2 are given. As indicated by the max-relevance and min-redundancy standard, the mRMR Score (mRMRS) of a_2 regarding a_1 has the additional structure:

$$mRMRS_{(a_2,a_1)} = MI(a_2,v) - MI(a_2,a_1) = (c_3 + c_4) - (c_3 + c_6) = c_4 - c_6 \quad (1)$$

where MI is observed between two factors,

$$\mathrm{CFS-SVM}_{(a_2,a_1)} = MIC(GSO(a_2,a_1),v) = (c_4 + c) - c_7 = c_4 \quad (2)$$

where $MI\ (a_2\ ,v)$ is the significance between the applicant feature a_2 and aimed variable t, $MI(a_2, a_1)$) is the repetition of a_2 as for already chosen feature a_1, and c_i indicates the i^{th} information area in Fig. 1. It can be seen that $MI\ (a_2\ ,a_1\ )$ includes two sections, c_3 and c_6. The two sections are considered as the important repetition and unimportant repetition individually, since a_3 denotes about the repetitive information that a_1 and a_2 conveys the purposes, however c_6 is totally irrelevant to the purpose. It has been planned that the unimportant repetition c_6 have to be removed from the mRMR score, leaving the main c_4, which denotes the special information about a_2 which transmits the target t [C.O. Sakar et al.,2012].

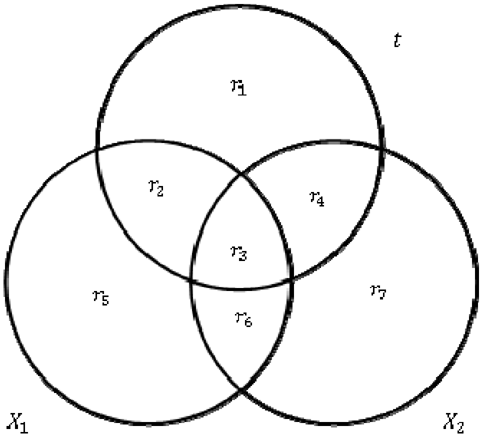


Fig.1 Relation Between Available Variable and Target Variable

In Fig.1. the association between three factors, including a variable v, a formerly chosen feature a_1 and a candidate feature a_2, where c_i shows the i^{th} information setting. Thus, a new channel feature selection method, CFS-SVM is designed to attempt to manage this issue. The Comprehensive Feature Selection based Support Vector Machine (CFS-SVM) value is given by:

where MIC (Maximal Information Coefficient) is the measurements between two factors of Gram-Schmidt Orthogonalization (GSO), that is between $GSO(a_2\ ,a_1\ )$ is the orthogonalized variable of candidate feature a_2 about early chosen feature a_1 by methods for GSO function and $MIC(GSO(a_2\ ,a\ ),v)$ where it shows the MIC significance between the orthogonalized variable $GSO(a_2\ ,a_1\ )$ and target variable t. In Fig.1, $GSO(a_2\ ,a_1)$ is denoted by $c_4 + c_7$, which is the information that is passed by a_2 however mainly from a_1 and $MIC(GSO(a_2\ ,a_1),v\ )$ can be viewed as subtracting c_7 from $c_4 + c_7$, just leaving c_4. It can be seen that the insignificant repetition c_6 is rejected from the CFS-SVM. In this way, the planned comprehensive technique can remove the insignificant repetition from the established mRMR.

3.2. Comprehensive Feature Selection (CFS) Method

An information network $X \in C^{D \times e}$ is tabled as D tests and E features, where i^{th} feature variable is $a_i \in C^D, 1 \leq i \leq E$. The objective variable $v \in Z^D$, with regards a different number denote about various classes of the compared samples. Beneath the high-relevance and low-redundancy standard, the proposed feature selection method CFS-SVM in the present test means to locate a least subset of helpful features $S \in C^{D \times e}, D < E$, where the i^{th} capable feature variable $s_i \in C^D, 1 \leq i \leq D$, and the objective variable v which could be perfectly described by the feature factors in S.

3.2.1. High-Relevance

The high-relevance plan is focused to choose the most helpful features, which effectively identifies special classes of tests. In the plan, the level of significance between the feature factors and target variable is measured, and the values of features are determined by the level of the significance score. To apply this plan, a recently investigated information metric MIC, was utilized as a part of the paper to score the significance level of a candidate feature variable as the objective variable. The High-Relevance Score (HRS) of feature variable a_i has the additional structure:

$$HRS_{(a_i)} = MINE_{MIC}(a_i, v) \quad (3)$$

where $MINE_{MIC}$ is a MIC score work from the Maximal Information-based Nonparametric study (MINE) application, which can be utilized to compute the MIC and other statistical scores. MIC is a statistical measure of the relationship between matched factors with regard to the relationship of linear or nonlinear. To receive the MIC score, the estimations of the two factors are divided into various numbers of cases to shape rectangular frames with various decisions. In this manner the delivery of the cells of every matrix can be received by giving the possibility chance to be grouped in every cell which is to be the part of focus falling into that particular cell, then the MI measurements for every network is determined and the high value is picked as the MIC score. The MIC of two factors a and v is characterized as:

$$MINE_{MIC}(a,v) = \max_{a_n v_n < D^{0.6}} \left[\frac{MI_{a_n, v_d}(a, v)}{\log_2 \min\{(a_d, v_d)\}} \right] \quad (4)$$

where D is the example measure a_n and v_n which indicates the quantity of cases forced on the a and v axis separately, and $MI_{a_d, v_d}(a, v)$ is the MI which observes the value between the two factors for an a_d -by - v_d rectangular network. As marked by the meaning of MIC, the MIC value can be utilized to calculate the level of significance between a constant variable and a individual variable, however the two existing MINE application packages independently, which are not a correct fit for getting a individual target variable as input, since the typical whole number estimations of target variable speak about various classes of the comparing tests, and these individual esteems will be confidently divided into cases whose number is generally not equivalent to the quantity of test classes.

Along these lines in the current research, the last bundle was conformed to guarantee that the estimation of target variable is just allowed to be merged into indicated holders whose number is dependable with that test classes. Then the significance between constant feature factors and the individual target variable was figured to utilize the balanced package with default parameters, where the more important HRS, the higher observable power of the relating feature.

3.2.2 High-Relevance (HR) and Low-Redundancy (LR)

The HR and LR expect to locate a minimal subset of informational features at the time of considering the subset. The basic solution of independently informational features does not really accomplish a better classification execution. Along these lines, "the *m* best features are not the best *m* features" [H. Peng et al.,2005]. Therefore, both the instruction of individual features and the duplication between them ought to be considered. There are various feature selection methods. For example, mRMR [H.

Peng et al.,2005] and Quadratic Programming Feature Selection (QPFS) [I.Rodriguez-Lujan et al.,2010], in which the distinction between the significance and redundancy is overestimated to improve the HR and LR.

In the present investigation, the HR and LR were in approximate way which was improved for utilizing an comprehensive technique which concentrates the GSO and MIC together. The GSO is dedicated for figuring the orthogonalized variable of a candidate feature as for other features to expel the redundancy amongst them, and the MIC significance between the orthogonalized variable and target variable is maximized in an approximate way which advances the high-relevance and low-redundancy.

The proposed OMICFS score of candidate feature variable a_i with respect to already chosen features $a_1, a_2, \dots, a_i$ is given by:

$$CFS-SVM_{(a_i, a_1, a_2, \dots a_j)} = MINE_{MIC}(GSO(a_i, a_1, a_2, \dots a_j), v) \quad (5)$$

where $GSO(a_i, a_1, a_2, \dots a_j)$ is the orthogonalized variable of feature a_i about the feature factors $a_1, a_2, \dots \dots, a_f$ utilizing a GSO work. The more notable CFS-SVM score, for all the more encouraging the feature a_i. By real, the unessential duplication is removed in CFS-SVM score in view of the input for orthogonal change and the target variable.

3.2.3 Implementation

In real time implementation, a step by step procedure is utilized to choose features that reach a nearby value of highly relevant maximized value. In the initial step, HRS (High Relevance Score) of the whole candidate features are calculated, where the feature holding high value of HRS can be determined as the primary trusted feature variable,

$$f_1 = arg \max_{a_i \in A} [HRS_{(a_i)}] \quad (6)$$

The comparing orthogonalized variable is chosen as $q_1 = \frac{f_1}{||f_1||}$. Along these lines, the issue turns out to be in the manner by which it incrementally chooses the other capable features from the rest of the features, where one feature denotes that one stage forward. Assuming that a feature subset $F_{z-1}, 2 \leq z \leq D$, made out of $m-1$ promising features $F_1, F_2, \dots \dots F_{z-1}$ which has been determined at step $z-1$ and the comparing orthogonalized factors are $q_1, q_2, \dots \dots q_{z-1}$. The m^{th} capable feature can be chosen from $mA - F_{z-1}$ at stage m by advancing the additional condition:

$$F_m = arg \max_{a_j \in (A - F_{z-1})} [CFS - SVM_{(a_i, F_{z-1})}]$$

$$F_m = arg \max_{a_j \in (A - F_{z-1})} [MINE_{MIC}(GSO(a_i, F_{z-1}), v] \quad (7)$$

where $GSO(a_i, F_{m-1})$ is the orthogonalized variable of candidate feature a_i as for the already chosen feature factors in F_{m-1} and it can be figured for utilizing a GSO work

$$GSO(a_i, S_{z-1}) = \frac{w_i}{||w_z||}, w_i = a - \frac{< a_i, g_i >}{< q_1, g_1 >} q_1 - \dots, - \frac{< a_i, g_{z-1}}{< q_{o-1}, g_{z-}} \quad (8)$$

With the assurance of F_o, the linked orthogonalized variable $q_m = GSO(F_o, F_{o-1})$ can be done parallel, which will be precious in following advances. Hence, the capable feature factors could be incrementally improved until step D where a sum of D features is chosen. For instance, there are 3 capable features that should be chosen from 5 capable features $X = \{a_1, a_2, a_3, a_4, a_5\}$.

- ✓ Regarding initial step, the HRS of the huge number of features is figured. Assuming a_2 has the maximal HRS, it can have $F_d = a_2, g_1 = \frac{a_2}{||a_2||}$.
- ✓ The second step starts with the calculation of $CFS - SVM_{(a_1, F_1)}, CFS - SVM_{(a_3, F_1)}, CFS - SVM_{(a_3, F_1)}, CFS - SVM_{(a_4, F_1)}, CFS - SVM_{(a_5, F_1)}$ and the feature a_3 , which is likely to

have the maximal OMICFS score which is chosen, leaving the other three features which to be picked later.

✓ In the third step, since $\mathrm{CFS-SVM}_{(a_1,F_1,F_2)}$ is considered as more important than $\mathrm{CFS-SVM}_{(a_4,F_1,F_2)}$ and $\mathrm{CFS-SVM}_{(a_3,F_1,F_2)}$. The last encouraging feature is determined as $F_3 = a_1, q_3 = GSO(F_3, F_1, F_2)$. In this manner, the three probable capable features $F = \{a_2, a_3, a_1\}$ are at last captured.

Moreover, the theory of Sure Independence Screening (SIS) [J. Fan and J. Lv.,2008] can be utilized to increase the speed of CFS-SVM when managing high-dimensional information. In each series of non-quickening CFS-SVM, the orthogonalized factors and MIC significance for every candidate feature which must be recalculated to be determine high CFS-SVM score. In this manner, when the feature measurement is highly-high, the computation becomes become tough. The SIS theory argues that the first arrangement of features can be pointed to a little subset whose measurement is in the request of $D/log\ D$ under the certain state of D, which provides the measurements of single features and size of tests individually, then feature selection can be a specialist by some advanced lower dimensional methods [J. Fan and J. Lv.,2008, Q. He and D.Y. Lin.,2011]. As per SIS, when OMICFS is utilized for high-dimensional and low-example estimate the data value, the candidate features can be positioned in dropping request by HRS, and just the best $\left[\frac{\lambda D}{logD}\right], \lambda \geq 1$ features are boosted into the stepwise search to reduce the computational weight.

The pseudo code of this proposed research work is given below:

```
Input: Candidate feature factors, $a_i \in A, 1 \leq i \leq E$ ,$A \in C^{D\times E}$ denoted as $D$ tests and $E$ features, and target variable $t \in Z^D$
Output: Ordinal number of D chosen features, D < E
for $i \leftarrow \{1,\dots,E\}$ do
    $MRelS_{(a_i)} = MINE_{MIC}(a_i, t)$
end for
if E >> D
    $A^c = \mathrm{rank}_{a_i \in A}[MRelS_{(a_i)}]$ in sliding request
    $A' = A^c_{(1,\dots,\left[\frac{\lambda D}{logD}\right],D)} < \left[\frac{\lambda D}{logD}\right] < E$
else
    $A' - A$
end if
$F_1 = \arg\max_{a'_i \in A'}[MRelS_{(a'_i)}]$
for $o \in \{2,\dots,D\}$ do
    $F_o = \arg\max_{a'_i \in A' - F_{o-1}}[MINE_{MIC}(GS$
    refresh $F = \{F_1, \dots\dots F_o\}$ 14: end for
restore the ordinal number of the chose features in $F$
```

3.3. Classifier

The Support Vector Machines (SVM) classifier is one among the machine learning algorithms which is based on the structural risk minimization principle and statistical learning theory. The basic idea of SVM is to transform the data into a higher dimensional space and find a classification hyper-plane that separates the data with the maximum margin. The standard SVM model is as follows:

$$f = \min\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i$$

Subject to (9)

$$y_i(\omega.\phi(x_i)) + b) \geq (1 - \xi_i)$$

where $\xi_i \geq 0, i = 1,2,\dots l$

where $x_i \in R^n$ and $y_i \in \{-1,+1\}$ are the training samples and the corresponding class label respectively, ϕ is a nonlinear map that transforms the data to the high dimensional feature space, ω is the normal vector to the bounding plane, b is a bias value, $\xi_i \geq 0, i = 1,2,\dots l$ are the slack variables, and C is a penalty parameter.

Instead of solving the above said optimization problem, it is feasible to solve the dual problem:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i,x_j) - \sum_{j=1}^{l}\alpha_j \qquad (10)$$

Subject to $\sum_{i=1} y_i\alpha_i = 0$

where $0 \leq \alpha_i \leq C, i = 1,2,\ldots,l$

where $k(x_i,x_j) = \phi(x_i).\ \phi(x_j)$ is called kernel function. The binary classifier $sign(g(x))$ and the decision function $f(x)$ could be calculated as follows:

$$f(x) = sign(g(x))$$

$$f(x) = sign\left(\sum_{x_i \in SV} y_i\alpha_i^* k(x,x_i) + b\right) \qquad (11)$$

where α_i^* is an optimal solution of the problem in eqn (10), $b = y_i \sum_{x_i \in SV} y_i\alpha_i^* k(x,x_i), if\ 0 < \alpha_i^* <$ C, SV is the set of support vectors,

4. ABOUT THE DATASET

The performance of the proposed CFS-SVM is experimented in PIMA Indian diabetes dataset [Kindie Biredagn Nahato et al.,2016] that has been obtained from National Institute of Diabetes and Digestive and Kidney Diseases. The chosen dataset is multivariate in nature and it contains 768 instances with 9 attributes including the class label. 268 instances are identified as the patients with gestational diabetes. The attribute contains mixture of integer and real numbers. The attribute information is shown in Table. 4.1.

Table 4.1. Attribute Information Of PIMA Dataset

S. No	Feature	Description	Domain	Zero Entry
1	Preg	Number of times pregnant	0-5	111
2	Glu	Plasma glucose concentration a 2 h in an oral glucose tolerance test	0-199	5
3	Bp	Diastolic blood pressure (mm Hg)	0-122	35
4	Skin	Triceps skin fold thickness (mm)	0-99	227
5	Insulin	2-h serum insulin (mu U/ml)	0-846	374
6	BMI	Body mass index (kg/mt sq)	0-67	11
7	DPF	Diabetes pedigree function	0.078-2.42	-
8	Age	Age (years)	21-81	-
9	Class	Class label	0 or 1	NIL

5. PERFORMANCE METRICS

Experiments were conducted on the selected clinical datasets using SCILAB 6.0.0. The performance metrics namely, accuracy, sensitivity, specificity, True Positive Rate (TPR), False Positive Rate (FPR) and precision were used for evaluating the proposed work. The metrics are computed by considering True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP). True positives (TP) refer to those instances that are truly identified as a diseased patient by the classifier. If the patient is not correctly classified, it becomes False Negatives (FN). Healthy instances correctly identified by the classifier becomes True Negatives (TN), if not it becomes False Positives (FP).

Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
True Positive Rate (TPR)	$\frac{TP}{TP+FN}$
False Positive Rate (FPR)	$\frac{FP}{TN+FP}$

Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$

6. RESULTS AND DISCUSSIONS

The overall performance analysis of the FELM and the proposed CFS based SVM (CFS-SVM) in terms of TP, TN, FP, FN, sensitivity and specificity is presented in Table-6.1. Also the performance analysis in terms of TPR, FPR, precision and Accuracy is portrayed in Table-6.2. As far as inferences from the results are concerned, the accuracy of the proposed CFS-SVM is improved and the time taken for feature selection is reduced. It is to be noted that the existing and proposed classifiers are allowed to train first and tested next.

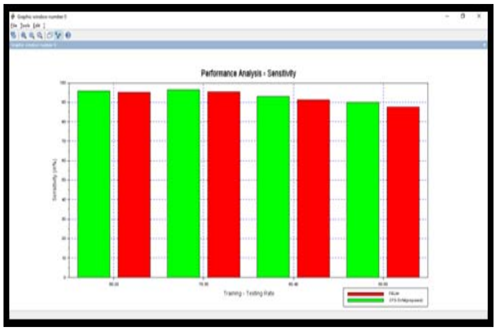


Fig.6.1 Performance Comparison Of Sensitivity For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

Sensitivity is the measure of proportion of actual positives which are correctly identified as positives by the classifier. From the Fig 6.1 it is clearly evident that the proposed classifier performs better in identifying the positives than the FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. The result values of Fig.6.1 are predicted in Table-6.1.

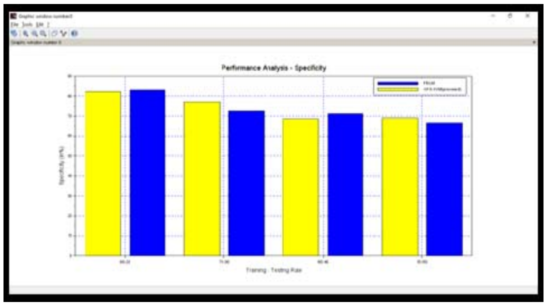
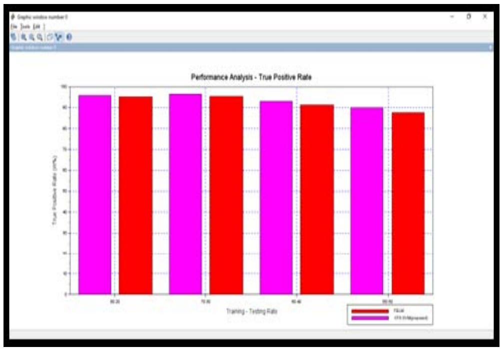


Fig.6.2 Performance Comparison Of Specificity For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

Specificity is the measure of classifier's ability to identify negative results. From the Fig. 6.2 it can be observed that the proposed mechanism does not work better in terms of specificity than the FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. This is due to the degree of relevance mismatch. The result values of Fig.6.2 are predicted in Table-6.1.



F

Fig.6.3 Performance Comparison Of True Positive Rate For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

True Positive Rate (TPR) refers to the positives that were correctly labelled by the classifier. From the Fig.6.3, it is evident that the proposed CFS-SVM produces better TPR than the FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. The result values of Fig.6.3 are predicted in Table-6.2.

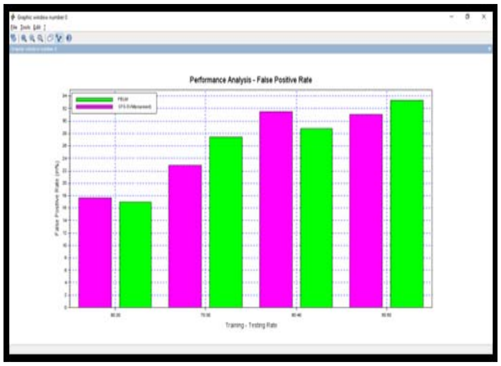


Fig.6.4 Performance Analysis Comparison Of False Positive Rate For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

False Positive Rate (FPR) refers to the negatives that were incorrectly labelled as positive. From the Fig. 6.4, it is clear that CFS-SVM attains the certain degree of relevance mismatch and results in producing a little bit of increase in the false positive rate in certain Training and Testing Rate when comparing with FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. The result values of Fig.6.4 are predicted in Table-6.2.

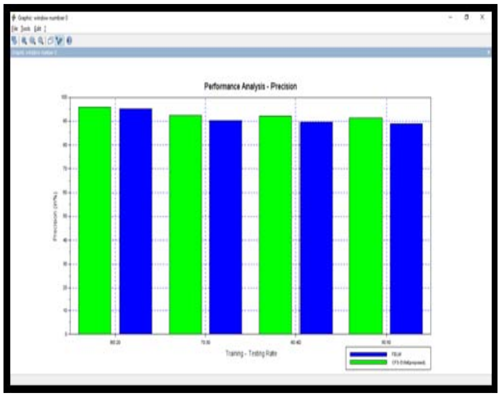


Fig.6.5 Performance Analysis Comparison Of Precision For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

Precision is the measure of accurately predicted positive values to the total predicted positive values. From Fig 6.5 it is clear that, CFS-SVM predicts the accurate positives than the FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. The result values of Fig.6.5 are predicted in Table-6.2.

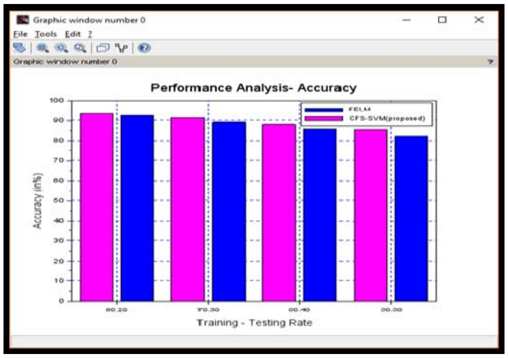
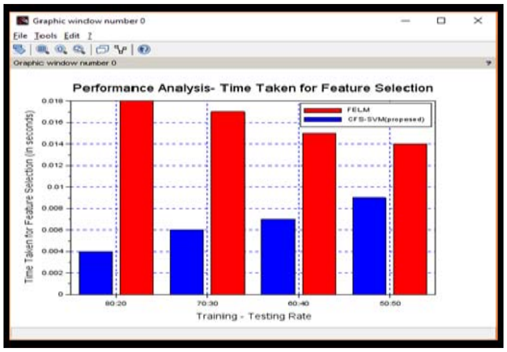


Fig.6.6 Performance Analysis Comparison Of Accuracy For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

Accuracy is the measure of ratio of correctly predicted observation to the total observations. The accuracy result is portrayed in Fig.6.7 and it illustrates that the proposed CFS-SVM harvests better accuracy than FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016]. The result values of Fig.6.6 are predicted in Table-6.2.



.

Fig.6.7 Performance Analysis Comparison Of Elapsed Time For PCA-SVM Vs CFS-SVM With Varying Training And Testing Rate

Time taken is the measure of finding how much period the algorithm takes for feature selection. Fig 6.7 shows that CFS-SVM took low time for feature selection than the FELM (Fuzzy-sets and Extreme Learning Machine) [Kindie Biredagn Nahato et al.,2016].

7. CONCLUSIONS

In this phase of research, comprehensive feature selection is applied for selecting the attributes from the dataset. The proposed feature selection mechanism employs maximum significance pattern in order to choose the most edifying features by which the degree of relevance between the feature variables and target variable is quantified, and the informativeness of features is determined according to the degree of the relevance score. A novel evidence metric, NEM, which is the statistical measure is used to score the relevance degree of a candidate feature variable with respect to the target variable. For obtaining the NEM score, the values of the two variables are partitioned into different number of bins to form rectangular grids with different resolutions, thus the distribution on the records in the dataset by letting the probability weight. After that certain statistics will be calculated, and the maximum is chosen as the MIC score. Once after the feature selection is carried out, SVM classifier is applied for performing the prediction of heart disease among gestational diabetes patients in the dataset. The results comparatively obtains overall better performance, but still there is a future scope of research in reducing the number of false positive and false negative. For this reason, in the next work we aim to employ an adaptive optimization technique for decreasing the false positive and false negative.

REFERENCES

[1] B. Dennis, S. Muthukrishnan, AGFS: Adaptive Genetic Fuzzy System for medical data classification, In Applied Soft Computing, Volume 25, 2014, Pages 242-252.

[2] B.M. Patil, R.C. Joshi, Durga Toshniwal, Hybrid prediction model for Type-2 diabetic patients, In Expert Systems with Applications, Volume 37, Issue 12, 2010, Pages 8102-8108.

[3] C.O. Sakar, O. Kursun, F. Gurgen, A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy–Maximum Relevance filter method, Expert Systems with Applications, 39 (2012) 3432-3437.

[4] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE transactions on pattern analysis and machine intelligence, 27 (2005) 1226-1238.

[5] I Kurt, M Ture, AT Kurum, Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, Expert Syst. Appl. 34 (1) (2008) 366–374.

[6] I. Rodriguez-Lujan, R. Huerta, C. Elkan, C.S. Cruz, Quadratic programming feature selection, Journal of Machine Learning Research, 11 (2010) 1491-1516.

[7] J Nahar, T Imam, KS Tickle, Y-PP Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Expert Syst. Appl. 40 (4) (2013) 1086–1093.

[8] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70 (2008) 849-911.

[9] J. Jabez Christopher, H. Khanna Nehemiah, A. Kannan, A Swarm Optimization approach for clinical knowledge mining, In Computer Methods and Programs in Biomedicine, Volume 121, Issue 3, 2015, Pages 137-148.

[10] J. Mattila, J. Koikkalainen, A. Virkki, M. van Gils and J. Lötjönen; for the Alzheimer's Disease Neuroimaging Initiative, "Design and Application of a Generic Clinical Decision Support System for Multiscale Data," in IEEE Transactions on Biomedical Engineering, vol. 59, no. 1, pp. 234-240, Jan. 2012.

[11] Jamal Salahaldeen Majeed Alneamy and Rahma Abdulwahid Hameed Alnaish, "Heart Disease Diagnosis Utilizing Hybrid Fuzzy Wavelet Neural Network and Teaching Learning Based Optimization Algorithm," Advances in Artificial Neural Systems, vol. 2014, Article ID 796323, 11 pages, 2014.

[12] Kalpana M, Kumar AS. Design and Implementation of Fuzzy Expert System using Fuzzy Assessment Methodology. Int J 2012;1(1):39–45.

[13] Kamadi V.S.R.P. Varma, Allam Appa Rao, T. Sita Maha Lakshmi, P.V. Nageswara Rao, A computational intelligence approach for a better diagnosis of diabetic patients, In Computers & Electrical Engineering, Volume 40, Issue 5, 2014, Pages 1758-1765.

[14] Kindie Biredagn Nahato, Khanna H. Nehemiah, A. Kannan, Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets, In Informatics in Medicine Unlocked, Volume 2, 2016, Pages 1-11.

[15] Kindie Biredagn Nahato, Khanna H. Nehemiah, A. Kannan, Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets, In

Informatics in Medicine Unlocked, Volume 2, 2016, Pages 1-11

[16] Manjeevan Seera, Chee Peng Lim, A hybrid intelligent system for medical data classification, In Expert Systems with Applications, Volume 41, Issue 5, 2014, Pages 2239-2249.

[17] MG Tsipouras, TP Exarchos, DI Fotiadis, AP Kotsia, KV Vakalis, KK Naka, et al., Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling, IEEE Trans. Inf. Technol. Biomed. 12 (4) (2008) 447–458.

[18] MJ Berry, G Linoff, Data Mining techniques: For marketing, sales, and Customer Support, John Wiley & Sons, Inc., 1997.

[19] Muhammad Waqar Aslam, Zhechen Zhu, Asoke Kumar Nandi, Feature Generation Using Genetic Programming With Comparative Partner Selection For Diabetes Classification, In Expert Systems With Applications, Volume 40, Issue 13, 2013, Pages 5402-5412.

[20] O. Kurşun, C.O. ŞAKAR, O. Favorov, N. Aydin, S.F. GÜRGEN, Using Covariates For Improving The Minimum Redundancy Maximum Relevance Feature Selection Method, Turkish Journal Of Electrical Engineering & Computer Sciences, 18 (2010) 975-989.

[21] Q. He, D.Y. Lin, A Variable Selection Method For Genome-Wide Association Studies, Bioinformatics, 27 (2011) 1-8.

[22] R Alizadehsani, J Habibi, MJ Hosseini, H Mashayekhi, R Boghrati, A Ghandeharioun, Et Al., A Data Mining Approach For Diagnosis Of Coronary Artery Disease, Comput. Methods Programs Biomed. 111 (1) (2013) 52–61.

[23] Vijaya K, Khanna NH, Kannan A, Bhuvaneswari NG. Fuzzy Neuro Genetic Approach For Predicting The Risk Of Cardiovascular Diseases. International Journal Of Data Mining, Modelling And Management 2010;2(4):388–402.

Table-6.1: Performance Analysis of TP, TN, FP, FN, Sensitivity and Specificity

Metrics →	TP		TN		FP		FN		Sensitivity		Specificity	
Algm / Rate	FELM	CFS-SVM	FELM	CFS-SVM	FELM	CFS-SVM	FELM	CFS-SVM	FELM	CFS-SVM	FELM	CFS-SVM
80-20	199	208	49	42	10	9	10	9	95.22	95.85	83.05	82.35
70-30	186	191	53	54	20	16	9	7	95.38	96.46	72.60	77.14
60-40	178	199	52	37	21	17	17	15	91.28	92.99	71.23	68.52
50-50	177	189	44	40	22	18	25	21	87.62	90.00	66.67	68.97

Table-6.2: Performance Analysis of TPR, FPR, Precision, F-1 Measure and Accuracy

Metrics →	TPR		FPR		Precision		Accuracy	
Algm / Rate	FELM	CFS-SVM	FELM	CFS-SVM	FELM	CFS-SVM	FELM	CFS-SVM
80-20	95.22	95.85	16.95	17.65	95.22	95.85	92.54	93.28
70-30	95.38	96.46	27.40	22.86	90.29	92.27	89.18	91.42
60-40	91.28	92.99	28.77	31.48	89.45	92.13	85.82	88.06
50-50	87.62	90.00	33.33	31.03	88.94	91.30	82.46	85.45