International Journal of Control and Automation
Vol. 12, No. 6, (2019), pp. 256-267

Grasshopper Optimization Algorithm based Feature Selection with Twin Support Vector Machine Classifier for Coronary Artery Heart Disease Prediction

[1]Rajkumar R,

[1]*Department of Computer Applications,Sri Krishna Arts and Science College, Coimbatore, India*

[2]Anandakumar K,

[2]*Department of Computer Applications,, Bannari Amman Institute of Technology,,Tamil Nadu, India*

[3]Bharathi A,

[3]*Department of IT,Bannari Amman Institute of Technology,Tamil Nadu, India*

Abstract

Data mining research extends its wings in almost all domains that include healthcare applications. ML algorithms took an important role in classification task and optimization algorithms are propounding in almost all tasks in knowledge discovery in data. This part of research work aims in employing grasshopper optimization algorithm on feature selection and twin SVM classifier is used for classification. This grasshopper optimization algorithm with twin SVM classifier is tested for performance over two dataset for prediction of CAHD. Performance metrics sensitivity, specificity, accuracy and elapsed time are considered formeasuring the effectiveness of the classifier. From the results it is inferred that the GOA-TSVM outperforms other chosen machine learning classifier.

Keywords: *Data mining, cardiac diseases, ML("Machine Learning"), SVM, Sensitivity , Specificity and Accuracy*

1. Introduction

Data mining is considered as the development of pulling out concealed knowledge from existing information which is capable to reveal the examples and connections among enormous measure of information in a solitary or a few datasets. Data mining is demonstrated to be actualized and working in numerous genuine applications which incorporate Risk recognition, evaluating risk and market investigation. An inclination of ventures like banking, finance, and advertising are utilizing Datamining for choping down expenses, and expand benefits. Cardiovascular diseases are among the most widely recognized reasons of death everywhere throughout the world. One significant sort of these maladies is coronary Artery heart disease (CAHD). Twenty five percent of individuals, who have CAHD, pass on all of a sudden with no past side effects. CAHD is one of the most basic sorts of illnesses influencing the heart, and potentially lead to extreme coronary attacks in patients. Monitoring the symptoms, can help in timely treatment, and decrease the seriousness of disease adverse effects. The motivation of the research work starts from these preliminaries.

The problem statement is quite obvious. Patients who have diabetes are more prone to CAHD. There are several ML algorithms and data mining algorithms are employed in this research arena. Many of the algorithmsdeal only with proposing a classifier. Only very few literatures are found that focuses on feature selection task before performing the classification task. Hence this research work concentrates on employing the efficient feature selection strategy by employing grasshopper optimization algorithm. After that,twin-SVM is used to perform the task of classification. The aim of this research is to reduce the elapsed time of the overall classification task during the testing phase. It is evident that without making use of feature selection strategy, the classifier will

ISSN: 2005-4297 IJCA

consume more time to perform the classification task. Hence there is a wide scope for employing an appropriate feature selection mechanism. This paper is structured as s. This section gives a quick view of the significance, motivation, problem statement. Section 2 discusses on the existing works. Section 3 portrays the proposed work. Section 4 narrates the dataset taken, performance metrics with results and discussions. Section 5 provides the concluding remarks of the paper.

2. Existing Works

Weighted Fuzzy Rule based Clinical Decision Support System [1] was proposed to automatically receive the information from the patient data. At first, weighted fuzzy rule generation approach was carried out to select and receive the attribute data. At last, DSS based on fuzzy centric logic was developed. The result of the comparison made with baseline schemes indicate that the systemhave low sensitivity and specificity.

Alphabet Entropy Method [2] was proposed to classify the cardiac arrhythmias which allow the inferences of prediction markers. With the use of nonlinear entropies, classification was done with the symbolic dynamics. Feature selection was given more priority and extraction was done with the help of random forest algorithm. In the result, F-Measure came with very low value.

Artificial Neural Network based Fuzzy System [3] was proposed to extract and make analysis to predict the heart failure. Variability of heart rate was used as base level signal, which was used as a input for the fuzzy system for the classification. This method was considered as feed-forward method, and result has low F-Measure.

Artificial Neural Networks based Decision System [4] was proposed with the assumption of heart rate attributes having common risk level. Deep analysis on the risk level provides the information that the artificial neural network won't work in a proper manner to detect and classify the status of heart rate, due to giving poor results on classification accuracy.

Enhanced Support Vector Machine Method [5] was proposed to classify the heart diseases among the patients. For the processing of input, Magneto Cardiograph signals were used. Actually it measures the magnetic fields produced from heart. Low Classification accuracy results indicate that the features were not fir for the prediction of heart disease.

Three Dimensional Cardiovascular System [6] was proposed with the base of images related to echo cardiographic. In order to construct the cardiovascular system, couple of algorithm was also proposed. To increase the interactivity, heart vessels thickness was considered as a feature. Low F-Measure shows that the system was not fit for the prediction of congenital heart disease.

Sophisticated Three Dimensional Classifier [7] was proposed to detect the diseases that are present in the heart vessels. This model was dependent on the patients precise pathological circumstance. Feature selection was performed to increase the accuracy, but the classification accuracy didn't got increased.

Data Quality Quantitative Assessment [8] was proposed to enhance the classification of high level frequency sounds from heart for increased classification towards prediction of coronary arteries. It is combined with the correlation analysis to provide an estimation towards the signal to noise ratio.

ISSN: 2005-4297 IJCA

Increased false positive rates shows that the prediction of coronary arteries is not possible only the heart beat sounds.

Sound Feature Selection [9] was proposed with the intention of classify the features based on heart sounds. The processing of digital signal was done with the help of data mining methods. Salient features which describes the low level frequency were identified. The low classification rate indicates that the method need further development to make more accurate results.

Dynamic Bayesian Network [10] was a temporal probability based graphic model, which focus on sequential events, its cause and dependency. It ensembles the concept of temporal abstractions for the risk level of Heart diseases. A network structure was built to study the parameters. The increased false negative rate indicates that there exist need to improve the method even more.

3. Proposed Work

The proposed work has two phases. In the initial phase, grasshopper optimization algorithm is employed for feature selection. In the second phase Twin-SVM classifier is used for performing the classification task. This is the extension of the previous work that can be found in the literature [15].

3.1. Feature Selection using Grasshopper Optimization Algorithm (GOA)

GOA starts improvement by making a lot of irregular arrangements. An arbitrary beginning populace framework of size is made and NV represents quantity of initial vector in populace and RNF was the necessary no of highlights be chosen. Every vector of the populace speaks to the lists of competitor highlights. Every component in previously mentioned vector is a whole number an incentive in 1 and the out number of highlights .hence lower limit of the hunt space of eacg measurement, alluded to by lb , is lb = 1, and upper limit, alluded by ub , is $ub = NoF$.

All up-and-comer operators are assessed with respect to a particular value of the fitness and the best search specialist in the ebb and flow populace is considered as objective. The blunder pace of characterization is considered as the fitness value. From that point onward, the hunt operators of the GOA update their positions and furthermore the diminishing component is acquired. During each cycle (emphasis), the situation of the best target acquired so far keeps on refreshed..In this AOGA, moreover the gotten qualities during the time spent reviving the circumstances of the grasshoppers are set then between. Thus, the range which is not in scope esteems are recognized , superseded against new and subjective entire numbers in the scope of $[1 - NoF]$. The other commitment GOA end of monotonous values in the acquired vectors. Another feature conveyance factor is included GOA to help the substitution of the copy features. By this additional upgrade in GOA, the probabilities on every component for instance subsets whose wellness esteem isn't actually the As necessities be, we portray the element factor of goodness for the component j in emphasis I as and compute it by

$$FG_{ji} = \alpha PG_j + (1 - PB_j)(NoF - DNF)/NoF \quad \ldots (1)$$

where PG_j represents probability that highlight j in subsets that are promising. is the likelihood of utilizing highlight j of less aggressive subsets is complete no of highlights and RNF is the necessary no of highlights to be chosen. is a positive consistent that stresses the significance of highlights in great subsets. In this work, this parameter is set to . To enlist the scattering probabilities of highlights,

take an instance of the essential component. Thus, PG will be 2/10, PB will be 4/10, and $FG = 2 \times 2/10 + (1-4/10)(6-3)/6 = 7/10$. At long last, the feature files are arranged by the most elevated FG value and its situation of FG , and after that the following higher value, etc. in a diving request.

Larger FG_j values demonstratesimproved highlightsthat are utilized for supplant the copy includes in the preliminary vector. The legitimization of the background thought is out of the way that the principal term in Eq. (1) demonstrates how much component j contributes in framing subsequent term shows the weighted likelihood that element j isn't chosen in less aggressive subsets. In this way, when modest no of highlights is required to be chosen, the job of the subsequent term is progressively conspicuous since this factor is near 1. Hence the features that are best will be obtained using GOA.Once GOA selects the features, then twin SVM are used in classification that are described in the next sub section.

3.2. Twin Support Vector Machine Classifier

TSVM is used in this research work for risk level classification of CAHD problem that loosens up the necessity that the hyperplanes are parallel in regular SVM, utilizing the condition (2)

$$f_1(x): w_1'x + b_1 = 0 \; and \; f_2(x): w_2'x + b_2 = 0 \ldots (2)$$

where w_1, w_2 and b_1, b_2 are common vectors in terms of bias in the equation of the cited on the couple of hyperplanes. As like conventional SVM it is presumed that the matrix $X_1 \in R^{m \times n}$ as the dat od label belong to "þ1" class, and $X_2 \in R^{m_2 \times n}$ since the data label belong to "+1", while $m_1 + m_2 = m$. Hence in order to acquire the above two proximal hyperplanes present in the equation (2), the optimization issues forthe TSVM will be coined as

$$\begin{aligned} &\min \frac{1}{2} \| X_1 w_1 + e_1 b_1 \| 2 + c_1 e'_2 \xi \\ &s.t \quad -(x_2 w_1 + e_2 b_1) + \xi \geq e_2, \qquad \xi \geq 0, \end{aligned} \quad \ldots (3)$$

and

$$\begin{aligned} &\min \frac{1}{2} \| X_2 w_2 + e_2 b_2 \| 2 + c_2 e'_1 \eta \\ &s.t \quad -(x_1 w_2 + e_1 b_2) + \eta \geq e_1, \qquad \eta \geq 0, \end{aligned} \quad \ldots (4)$$

where c1and c2 are the punishment mode parameters, and ξ,η are the leeway type of vectors. It very well may be seen that the first term in the target work present in the condition (10) is utilized to create "+1" marked events that are proximatedwith the hyperplane $w'_1 x + b_1 = 0$, where the 2nd foace and constraints labels are bonded in hyperplane $w'_1 x + b_1 = -1$. In order to obtain the solutions of problems (3) and (4), it is derived the CAHD classification as dual problems as

$$\begin{aligned} &\max_{\alpha} e'_2 \alpha - \frac{1}{2} \alpha' G (H'H)^{-1} G' \alpha \\ &s.t \; 0 \leq \alpha \leq c_1 e_2. \end{aligned} \quad (5)$$

and

$$\max_{\beta} e'_1 \beta - \frac{1}{2}\beta' H(G'G)^{-1} H'\beta \qquad (6)$$
$$s.t\ 0 \le \beta \le c_1 e_1.$$

where H=[A e_1] ,G=[B e_2] and j=[Xe]. By observing (5) and (6), it can be observed that TSVM provides a solution to the couple of small size instead of a single large which is present in the conventional support vector machine. When the solutions such as , α and β of the problem in (5) and (13) are concieved, the non-parallel and proximal level of hyperplanes (2) are obtained by

$$\begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = -(HH)^{-1} G'\alpha \; and \; \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = -(GG)^{-1} H'\beta \qquad (7)$$

A fresh and new instances of X is then assigned for the label " +1" and "-1" which depends on the proximal level of the hyperplanes as in (7) and it lies cloaser. By measuring the hyperplanes, it shows that TSVM is capable enough to classify patients those who are prone to CAHD.

4. Datasets and Performance Metrics

Two datasets namely Statlog Heart dataset [16] and UCI heart disease PIMA dataset [17] are taken for performance evaluation. Statlog heart dataset contains 270 instances with 14 attributes including class label. Out of the 270 instances 120 instances are true positive and remaining 150 instances are true negative.PIMA dataset contains 768 instances with 9 attributes including class label. Out of the 768 instances 268 instances are true positive and remaining 500 instances are true negative. The details of the dataset are given in chart format in table 1. The metrics that are used for the performance measures are TP ("True Positive"), TN ("True Negative") , and similarly false values namely FP and FN. The other metrics include accuracy, specificity, sensitivity and elapsed time and are compared with other classifiers available.

Table – 1. Dataset Details

Statlog dataset		**PIMA dataset**	
Total instances	270	**Total instances**	768
Number of attributes	14	**Number of attributes**	9
Attributes	1) Age 2) Male / Female 3) Type of pain 4) BP while rest 5) Cholestrol in Serum 6) FBS 7) electrocardiographic result in rest 8) maximum pulse 9) Agina for excersice 10) Old-peak 11) Major vessel count	**Attributes**	1) Pregnancy count 2) Glucose level concentration 3) BP in diastolic 4) Thickness of skin fold in triceps 5) 2-Hour serum insulin 6) BMI 7) DPF 8) Class label

	12) Value of thalium 13) Class label		
TP	120	**TP**	268
TN	150	**TN**	500

5. Results and Discussions

Table – 2. Performance Evaluation – Statlog Dataset

Method	TP	TN	FP	FN	Accuracy (in %)	Sensitivity (in %)	Specificity (in %)	Elapsed Time (in seconds)
Firefly Algorithm[11]	104	122	21	23	83.70	81.89	85.31	497.248
Neural Networks Classifier [12]	112	126	16	16	88.15	87.50	88.73	361.881
Modified Differential Evolution with Neural Networks [13]	113	127	15	15	88.89	88.28	89.44	318.765
Neuro Fuzzy Classifier [14]	116	133	11	10	92.22	92.06	92.36	223.324
TSVM based ImprocvedBacterial Forageing[15]	117	136	9	8	93.70	93.60	93.79	38.475
Twin SVM with GOA	118	142	6	4	96.30	96.72	95.95	21.164

Table – 3. Performance Evaluation – PIMA Dataset

Method	TP	TN	FP	FN	Accuracy (in %)	Sensitivity (in %)	Specificity (in %)	Elapsed Time (in seconds)
Firefly Algorithm[11]	207	435	61	65	83.59	76.10	87.70	1527.781
Neural Networks Classifier [12]	231	439	48	50	87.24	82.21	90.14	1304.362
Modified Differential Evolution with Neural Networks [13]	235	441	44	48	88.02	83.04	90.93	1086.115
Neuro Fuzzy Classifier [14]	242	450	37	39	90.10	86.12	92.40	697.164
Improved Bacterial	256	455	30	27	92.58	90.46	93.81	108.574

Foraging Optimization based Twin Support Vector Machine [15]								
Twin SVM with GOA	264	468	20	16	95.31	94.29	95.90	51.875

Fig.1. shows the performance of the classifiers in terms of accuracy for the statlog dataset. It is understood from the results that the proposed TSVM with GOA obtains 96.30% accuracy which is better than that of all classifiers. It is because of the improved true positive and true negative values as well as reduced false positive and false negative values. The same also impacts on sensitivity which is obtained 96.72 % (shown in Fig.2) and specificity obtained 95.95% (shown in Fig.3). The total elapsed time is eventually reduced to 21.164 seconds (shown in Fig.4) which is significantly less when compared to other classifiers.

Fig.5. shows the performance of the classifiers in terms of accuracy for the PIMA datasetIt is understood from the resultsthat the proposed TSVM with GOA obtains 95.31% accuracy which is better than that of all classifiers. It is because of the improved true positive and true negative values as well as reduced false positive and false negative values. The same also impacts on sensitivity which is obtained 94.29 % (shown in Fig.6) and specificity obtained 95.90% (shown in Fig.7). The total elapsed time is eventually reduced to 51.875 seconds (shown in Fig.8) which is significantly less when compared to other classifiers.

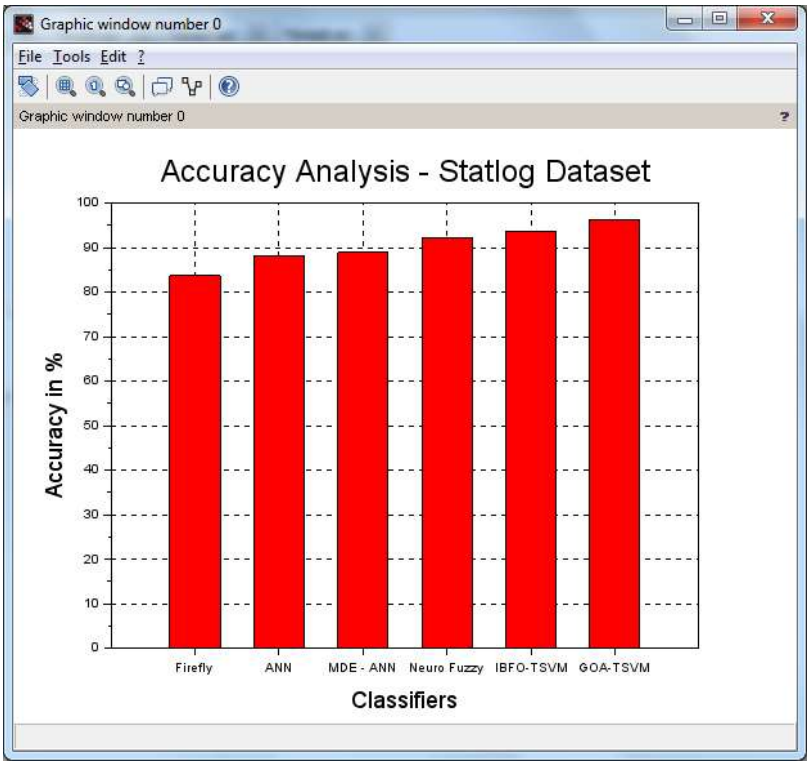

Fig.1. Accuracy Analysis

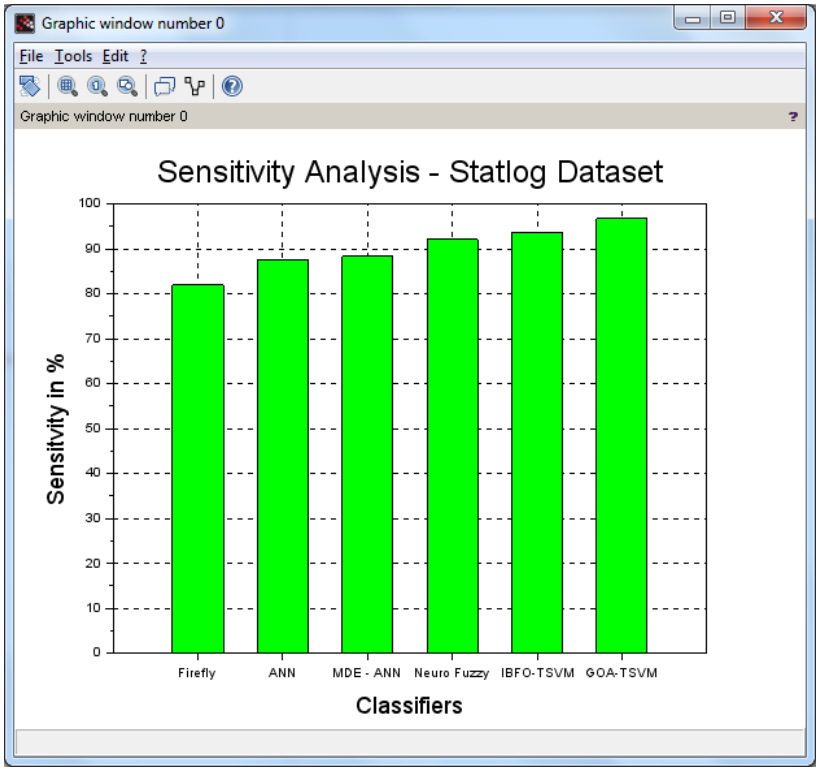

Fig.2. Sensitivity Analysis

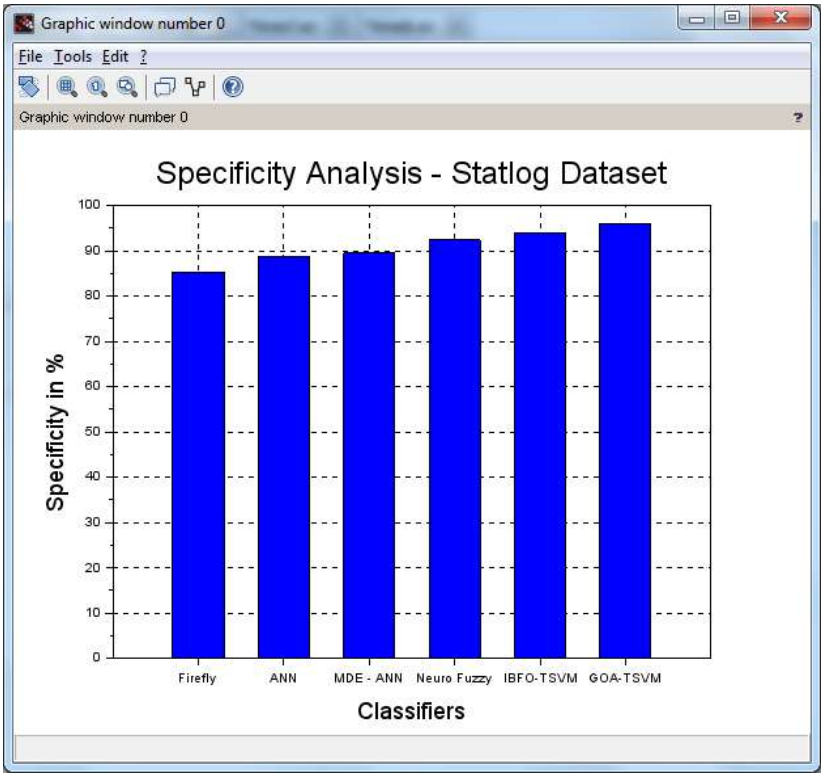

Fig.3. Specificity Analysis

ISSN: 2005-4297 IJCA

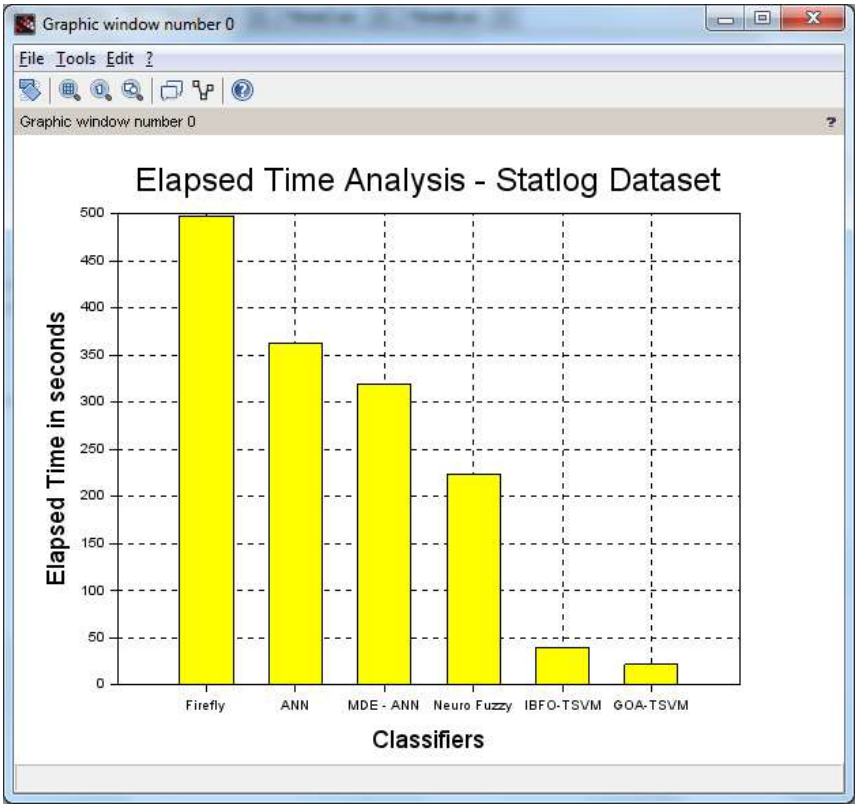

Fig.4. Elapsed Time Analysis

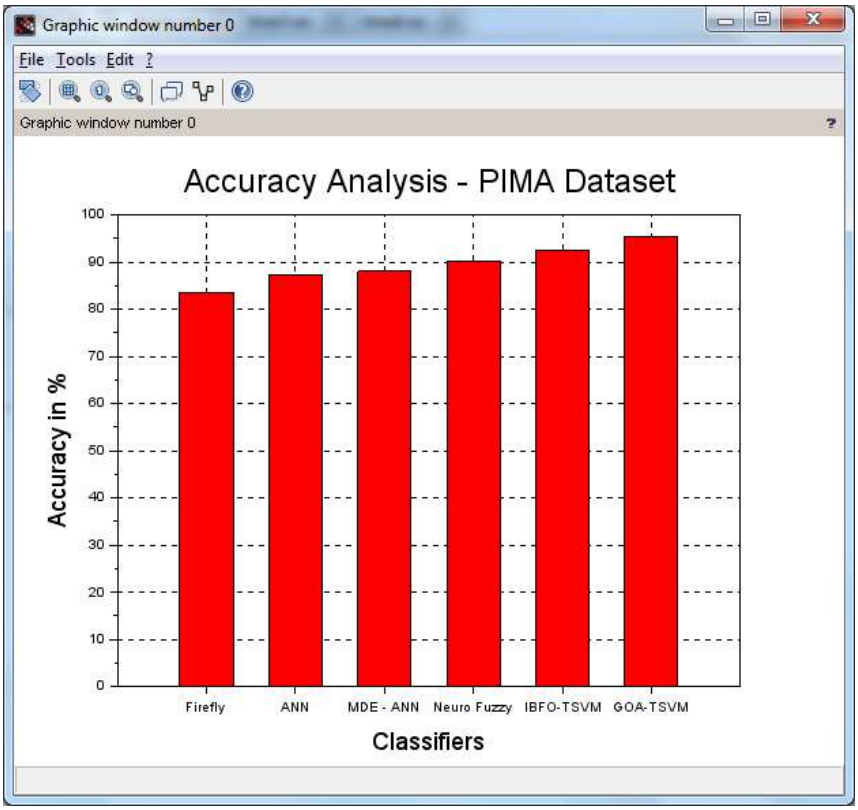

Fig.5. Accuracy Analysis

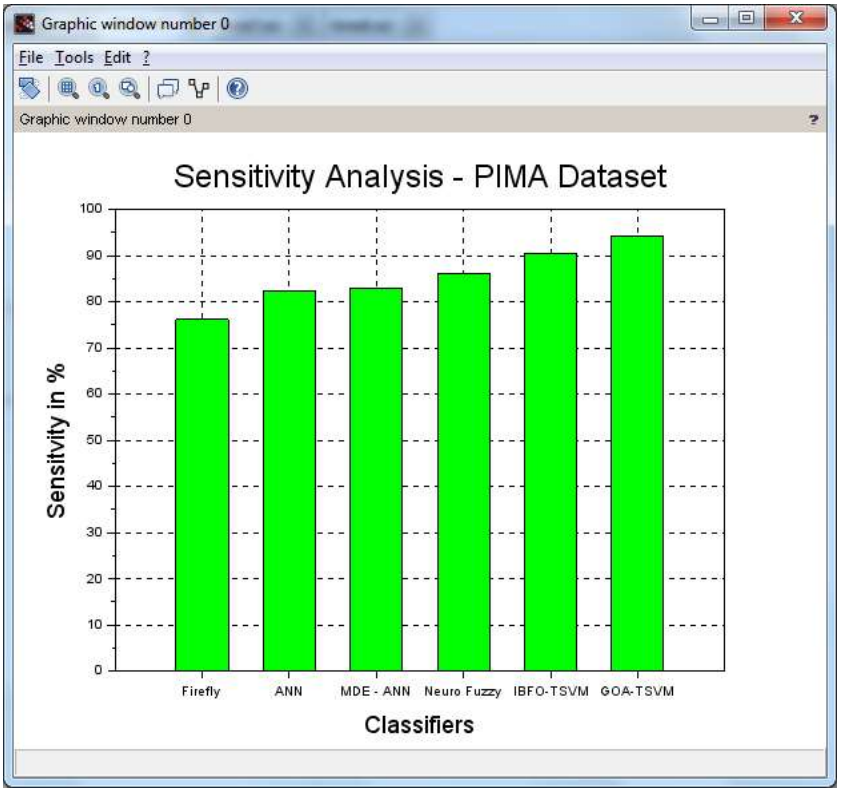

Fig.6. Sensitivity Analysis

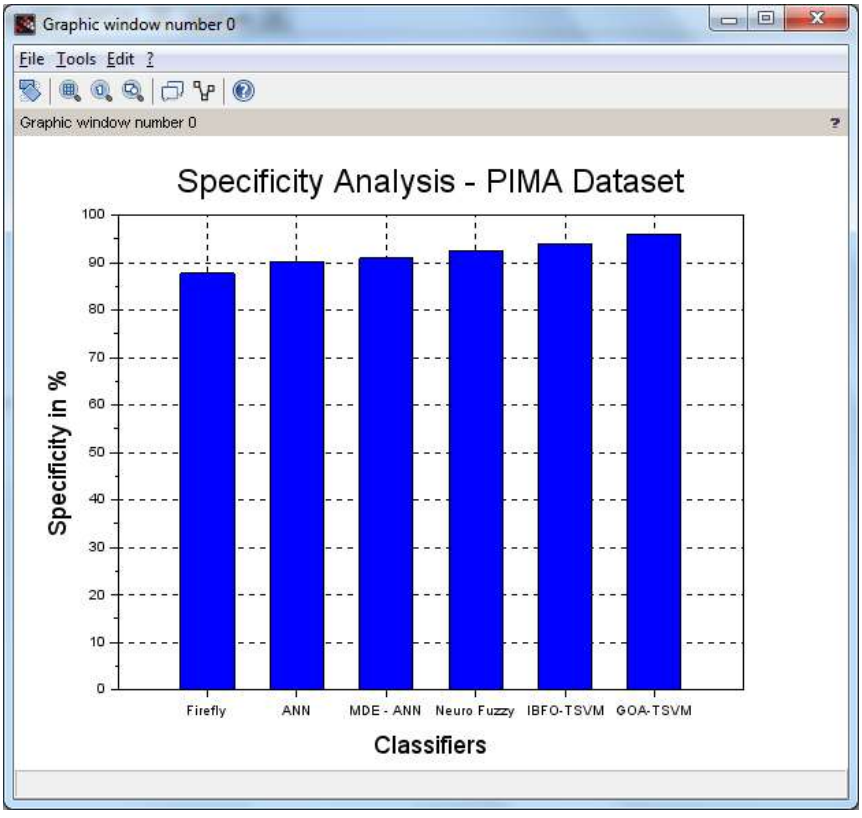

Fig.7. Specificity Analysis

ISSN: 2005-4297 IJCA

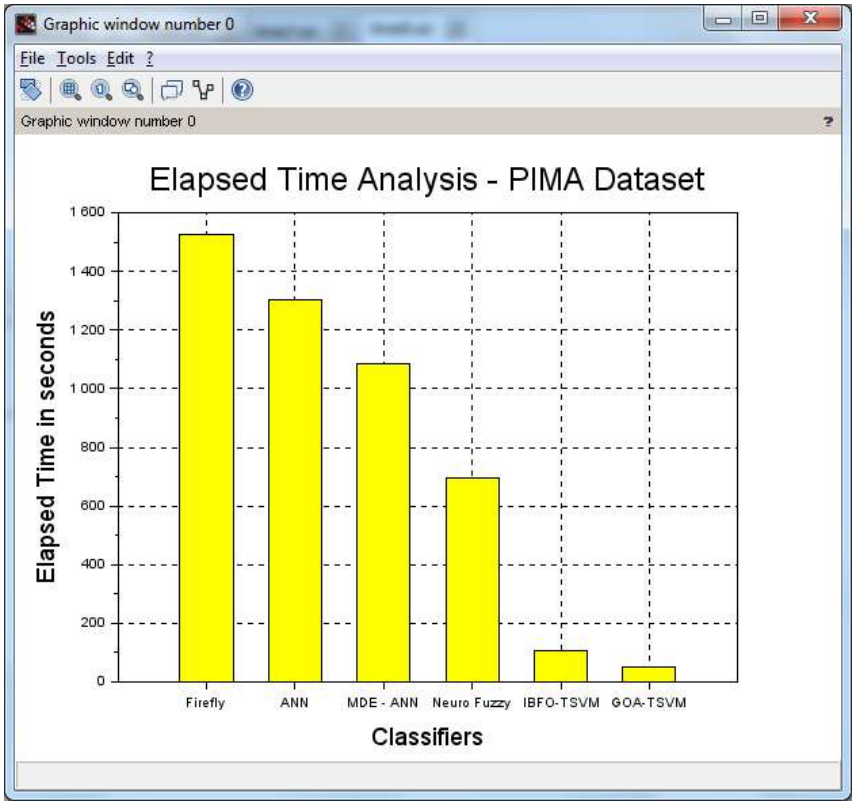

Fig.8. Elapsed Time Analysis

6. Conclusion

Twin SVM classifier is enhanced in this research work and grasshopper optimization is employed for performing the feature selection task. The reduced feature set is given as the input for the TSVM classifier. This work is the extension of the previous work done namely Twin SVM for Improved Bacterial Foraging Optimization. From the obtained results it is understood that TSVM with GOA feature selection outperforms than that of existing chosen classifiers in terms of selected performance metrics.

References

[1] A. J. Amutha, R. Padmajavalli, D. Prabhakar, A novel approach for the prediction of treadmill test in cardiology using data mining algorithms implemented as a mobile application, Indian Heart Journal, Vol 70, Pages 511-518, 2018.

[2] M. Tayefi, M. Tajfard, S. Saffar, P. Hanachi, A. R. Amirabadizadeh, H. Esmaeily, A. Taghipour, G. A. Ferns, M. Moohebati, M. Ghayour-Mobarhan, hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm, Computer Methods and Programs in Biomedicine, Vol 141, pp 105-109, 2017.

[3] T. Vivekanandan, N. C. Iyengar, Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Computers in Biology and Medicine, Vol 90, pp 125-136, 2017.

[4] M. S. Amin, Y. K. Chiam, K. D. Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, Vol 36, pp 82-93, 2019.

[5] L. Wang, P. J. Haug, G. D. Fiol, Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository, Journal of Biomedical Informatics, Vol 69, pp 259-266, 2017.

ISSN: 2005-4297 IJCA

[6] S. P. Potharaju, M. Sreedevi, V. K. Ande, R. K. Tirandasu, Data mining approach for accelerating the classification accuracy of cardiotocography, Clinical Epidemiology and Global Health, 2018.
[7] M. Nilashi, O. b. Ibrahim, H. Ahmadi, L. Shahmoradi, An analytical method for diseases prediction using machine learning techniques, Computers & Chemical Engineering, Vol 106, pp 212-223, 2017.
[8] M. A. jabbar, B. L. Deekshatulu, P. Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, Procedia Technology, Vol 10, pp 85-94, 2013.
[9] M. Ilayaraja, T. Meyyappan, Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets, Procedia Computer Science, Vol 70, ppj 586-592, 2015.
[10] N. M. Rémy, T. T. Martial, T. D. Clémentin, The prediction of good physicians for prospective diagnosis using data mining, Informatics in Medicine Unlocked, Vol 12, pp 120-127, 2018.
[11] N. C. Long, P. Meesad, H. Unger, "A Highly Accurate Firefly Based Algorithm for Heart Disease Prediction", Expert Systems with Applications, vol. 42, no. 21, pp. 8221 – 8231, 2015.
[12] C. H. Weng, T. C. K. Huang, R. P. Han, "Disease prediction with different types of neural network classifiers", Telematics and Informatics, vol. 33, no. 2, pp. 277 – 292, 2016.
[13] T. Vivekanandan, N. ChSrimanNarayanaIyengar, "Optimal Feature Selection using a Modified Differential Evolution Algorithm and its Effectiveness for Prediction of Heart Disease", Computers in Biology and Medicine, vol. 90, pp. 125 – 136, 2017.
[14] R. Rajkumar, K. Anandakumar, A. Bharathi, "Risk Level Classification of Coronary Artery Heart Disease in Diabetic Patients using Neuro Fuzzy Classifier", International Journal of Computational Intelligence Research, vol. 13, no. 4, pp. 575 – 582, 2017.
[15] R. Rajkumar, K. Anandakumar, A. Bharathi, "Improved Bacterial Foraging Optimization based Twin Support Vector Machine (IBFO-TSVM) Classifier for Risk Level Classification of Coronary Artery Heart Disease in Diabetic Patients", International Journal of Applied Engineering Research, vol. 13, no. 3, pp. 1716 – 1721, 2018.
[16] http://archive.ics.uci.edu/ml/datasets/statlog+(heart)
[17] https://archive.ics.uci.edu/ml/datasets/diabetes

ISSN: 2005-4297 IJCA