International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-8, June 2020

# Evolutionary Multivariate Kernal Svm Prediction Method for Classification

**K.Geetha**

***ABSTRACT : Thyroid disorders are common among the world wide population. This disorders posses' significant problems among Indians. Research studies shows that nearly 32% of Indian population suffers from various thyroid disorders. This paper deals with the thyroid data set which in turn classify into three groups as hyper thyroidisim, hypothyroidism and normal. The American Thyroid Association reported twelve percent of their citizens suffer from thyroidism in which 60% population are unaware of their condtions.. Above statistics implies the classification of thyroid disorder is crucial in global perspective too. The thyroid data set are collected from UCI repository and it is multivariate type with 21 attributes. With the 21 attributes only 10 attributes are selected based on their rank. Hybrid Differential Evolution Kernel Based SVM algorithm is used to classify the data set. It takes around 30 epochs to stabilize the errors. The classification accurancy is observed to be 67.97%.***

***Keywords: Curse of Dimensionality , Classification ,Evolutionary algorithm , Multivariate data type ,Thyroid Data Set***

## I. INTRODUCTION

According to the statistics published by the WEEK magazine in India, it reported that the percentage of people suffered from throid disorders are in rise. The above report based on the results derieved from survey conducted by Indian Thyroid Society. The statistics shows that women of age between 18 to 35 are more prone to thyroid disorders compared to men. The thyroid disease is ranked nineth in comparison to other syndromes like diabetic, Blood Pressure, respiratory Problems, insomania and heart related issues. A statistical survey revealed that fifty percent of the surveid population are aware of thyroid disorder; knows about diagnostic test and remedies for this health issues. Other research work on medical data set depicts , C.V.Subbulakshmi et.al proposed hybrid algorithm self-regulated particle swarm optimization (SRPSO) algorithm with the extreme learning machine (ELM) for classification problems.To optimize the input weights and hidden biases and minimum norm least-square scheme, an improved PSO is used to analytically determine the output weights[1]. Tapas Ranjan and Subhendu Kumar's research work use decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm to classify these diseases and compare the effectiveness, correction rate among them. Performance of the mentioned classifiers shows Multilayers Perceptron gives oves all best classification result whereas Naïve Bayes performance is not significant [2]

**Revised Manuscript Received on June 30, 2020.**
***** Correspondence Author
**Dr. K.Geetha*,** Department of Computer Applications, Sri Krishna Adthiya College of Arts and Science, Coimbatore, Tamilnadu, India. Email: kkgeetha17@gmail.com



This research work classifies the thyroid disorder with the help of proper features selection from the dataset which in turn leads to accurate results which helps physician in diagnosis of thyroid.

## II. METHODOLOGY

To classify the data into three classes like Hypothyroid,Hyperthyroid and Normal , get the dataset from any repository or UCI repository , in this nearly eight thousand data set with twenty one attributes are taken into account in which fifteen attributes have continous values and six are discreete type. First it undergoes a preprocessing stage that is followed by feature selection . It is carried out to overcome the curse of dimensionality issues using hybrid algorithm Differential evolution with Support Vector Machine. The subset of the data set is classified using kernal based SupportVector Machine algorithm based on the fitness calculation where error stabilization is used to measure the fitness.

### 2.1 Preprocessing Step

The accuracy of the results greatly depend on the type of data set , which is free from the problems like redundant data, missing values and noisy data etc. In this research Not a Number and missing value constraints are checked .Data set contains both the types of values such as continous and discrete.

### 2.2 Feature Selection Method

Challenges in using medical database are its volume,heterogenous values and privacy issues. Subset selection methods or ranking methods are used in feature selection to identify the relevant attributes .This step helps to overcome curse of dimensionality issues by removing irrelevant and redundant data, improve the quality of data and its accuracy. The algorithms used for Feature selection are classified as

1)The filter model 2) The wrapper model

The wrapper Model gives good accuracy but more complex computational procedures have to be carried out. The proper classification depends on the selection of data. The above process leads to over fitting of data. Various wrapper methods are available for each domain.

### 2.3 Differential Evolution

Differential Evolution is an effective search method for continuous optimization problems. The steps involved are mutation, crossover and selection.

The process of differential evolution differs from other evolutionary algorithm is the mutation that perturbs the selected vector according to the scaled difference of the other two members of the population. The DE algorithm:

1: Initialize the parameters
2: Set the termination condition
3: Loop the population member vector $p_i$ do
4: Mutant vector $m_i$ is created
5.The population member $p_i$ and mutant vector is crossover and create trial vector $tr_i$
6: end Loop
7: Loop the population member vector $p_i$ do
8: Check fitness of trial vector less than fitness of population vector then
9: Assign trial vector as population vector
10: repeat the steps 8 and 9 until termination condition reached

The NP represent population size and each vector $v_i$, of dimensionality D, consists of real-valued parameters, $v_i = (v^1_i , . . . , v^D_i) \in RD$, for i = 1, . . .,NP. Usually the population is initialized with vectors of values obtained randomly in the interval $[v_{lb}, v_{ub}]$, where $v_{lb}$ and $v_{ub}$ signifies the lower and upper bound, of the vector. New population is generated through mutation and crossover in each iteration. The trial vectors stores the new population. During mutation,a new mutant vector is formed for each member of target vector or current population members. The mutation is happened according to

$$\mathbf{m}_i = \mathbf{P}r1 + F \cdot (\mathbf{P}_{r2} - \mathbf{P}_{r3})$$

where $m_i$ is a mutant while Pr1, Pr2 and Pr3 are population vectors which are selected at random at the condition $i \neq r1 \neq r2 \neq r3$, and $F \in [0,\infty)$. Mutation occurs after crossover which happens between the target vector and the corresponding mutant vector creating a trial vector. The crossover is done as follows:

$$tr_{ji} = \begin{cases} m^j_i & \text{if } m[0, 1) \leq cr \text{ or } j = r_j , \\ P^j_i & \text{otherwise} \end{cases}$$

where j = 1, . . .,D. The above crossover is called the binomial crossover. After the trial vector is obtained , which is treated as new population and move to next generation. The fitness function plays a major role in selecting the population by replacing the target vector with trail vector if the fitness fuction is lesser or equal to given objective specified in the fitness function. Due to its simplicity and fastness differential evolutionary algorithms are mostly applied in optimization problems.

## III. CLASSIFICATION

The function F of the classifier maps the feature vector in input to the labels in output class i ∈ I where I is the feature space. Its easier to compute kernal function but its hard to compute the feature vector corresponding to the kernel. In RBF kernel ( k(x,y) = exp( -||x-y||^2) the corresponding feature vector is infinite dimensional.

### 3.1 Kernel Trick

Many machine learning algorithms can be written use dot products, and then dot products replaced by kernels. By doing so, it is not necessary to use the feature vector at all and also helps in complex computations in highly performing kernals in efficient way without writing down the huge and potentially infinite dimensional feature vector. The kernal trick emphasis on the probable to stuck with relatively low dimensional, low-performance feature vectors when its not possible to work with kernal functions directly.

The idea behind the kernel trick is to map the data in to a feature space and construct a linear classifier. In this space, non-linear classifiers in the original space can be constructed.

### 3. 2 Radial Basis Kernal Function

In this work, data has high dimensional feature and it is non-linear. In order to make the data linear and reduce the dimensionality (Curse of Dimensionality), RBF kernel is used. The RBF kernel on two samples x and x', represented as feature vectors in some input space, is defined as[3]

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

$||\mathbf{x} - \mathbf{x}'||^2$ is recognized as the squared Euclidean distance between the two feature vectors. $\sigma$ is a free parameter. An equivalent, but simpler, definition involves a parameter $\gamma = \frac{1}{2\sigma^2}$:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2)$$

Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one (when x = x'), it has a ready interpretation as a similarity measure[5]. The feature space of the kernel has an infinite number of dimensions; for $\sigma = 1$, its expansion is [4]:

$$\exp\left(-\frac{1}{2}||\mathbf{x} - \mathbf{x}'||^2\right) = \sum_{j=0}^{\infty} \frac{(\mathbf{x}^\top \mathbf{x}')^j}{j!} \exp\left(-\frac{1}{2}||\mathbf{x}||^2\right) \exp\left(-\frac{1}{2}||\mathbf{x}'||^2\right)$$

## IV. RESULTS AND DISCUSSION

Matlab is used to carried out this work. The figure 1 shows the framework of the model where the dataset is loaded. Figure 2 and 3 displays the preprocessed data set and the list of features selected , the same figure also represent the error stabilization criteria..

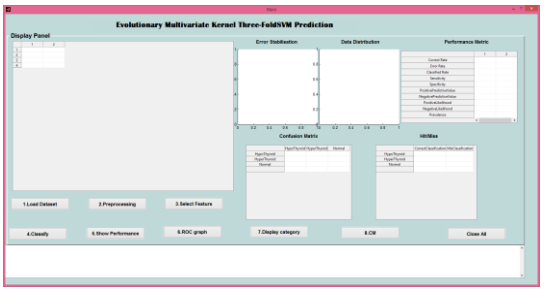


**Fig 1: Load the Data**

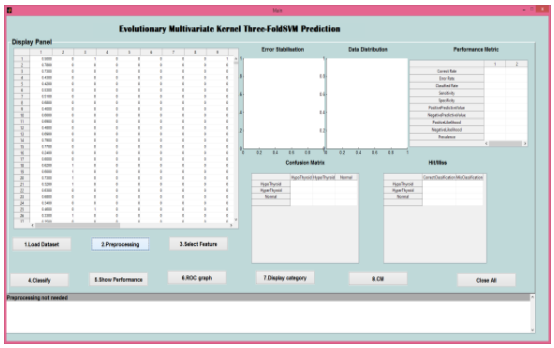


**Fig 2: List the Preprocessed Data Set**

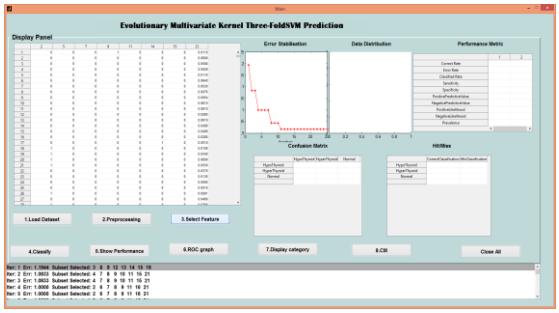


**Fig 3: List the Selected Features and Error Stabilization**

### 4.1 Performance Metrics

The evaluation of the classification is essential. The performance metrics that are included in this methodology is elaborated as follows and the output is shown in figure 4. The results are depicted in confusion matrix to differetiate the true and false situation of the medical conditions. The metric prevalence (Pr) for the considered group.

$$\Pr = \frac{TP + FN}{TP + FN + TN + FP}$$

***Table 2: Confusion Matirx***

| | | Test result | |
|---|---|---|---|
| | | 0 | 1 |
| True situation | 0 | True negative (TN) | False positive (FP) |
| | 1 | False negative (FN) | True positive (TP) |

The metric sensitivity (Se), specificity (Sp). Calculated as

$$Se = \frac{TP}{TP + FN} \qquad SP = \frac{TN}{TN + FP}$$

The accuracy is measured by

$$ACC = \Pr * Se + (1 - \Pr) * SP$$

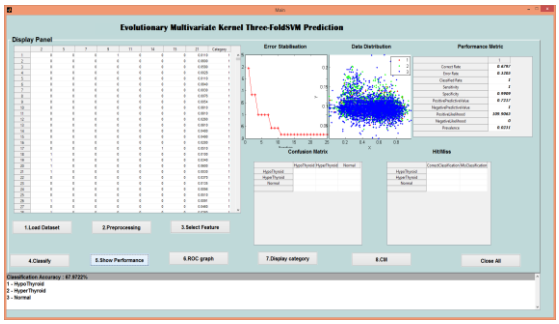


**Fig 4: Classification and Evalaution metrics**

It takes nearly thirty epohs to error stabilization Then data is classified into three class labelsas depict in figure 4. 67.97% accuracy level is reached in this classification. ROC for the method is shown in figure 5.

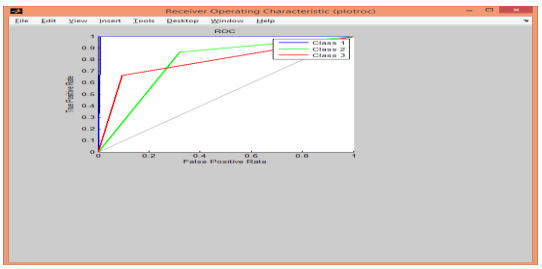


**Fig 5: Receiver Operating Characteristic (ROC)**

## V. CONCLUSION

The suggested work classify data into respective class labels of thyroid. The evaluation metrics are used to evaluate the accuracy which shows as 67.97%. It doesn't show better result as compared with other hybrid models like differntial evolution with naïve bayes. The processing time for this model is higher and it takes nearly thirty runs to compute the fitness for eight thousand data. The future work can be test with genetic algorithm with SVM for multivariate data set.

## REFERENCES:

1. C. V. Subbulakshmi and S. N. Deepa, "Medical dataset classification: a machine learning paradigm integrating particle swarm optimization with extreme learning machine classifier," The Scientific World Journal, vol. 2015, Article ID 418060, 12 pages, 2015.
2. Tapas Ranjan Baitharu & Subhendu Kumar Pani,"Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset", Procedia Computer Science
3. Volume 85, Pages 862-870, 2016.
4. Vert, Jean-Philippe, Koji Tsuda, and Bernhard Schölkopf (2004). "A primer on kernel methods".
5. Shashua, Amnon (2009). "Introduction to Machine Learning: Class Notes 67577".

## AUTHOR PROFILE

**Dr. K.Geetha,** Ph.D, is a Professor and Researcher at the Department of Computer Applications, Sri Krishna Adthiya College of Arts and Science,Coimbatore, Tamilnadu, India. She is having 16 years of teaching experience and 8 years of research experience. She published 12 research papers in Web of Science,Scopus Indexed and peer reviewed journals. She presented 25 research articles in various national and international conferences. Her research articles mostly focused on bio-inspired computing. She is fond of introducing new technologies in teaching learning process. She is currently doing research in technologies in teaching learning process.