Global Journal of Computer Science and Technology: E Network, Web & Security
Volume 16 Issue 1 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method

By K.Geetha & Capt. S. Santhosh Baboo
Periyar University, India

Abstract- Thyroid diseases are widespread worldwide. In India too, there is a significant problems caused due to thyroid diseases. Various research studies estimates that about 42 million people in India suffer from thyroid diseases [4]. There are a number of possible thyroid diseases and disorders, including thyroiditis and thyroid cancer. This paper focuses on the classification of two of the most common thyroid disorders are hyperthyroidism and hypothyroidism among the public. The National Institutes of Health (NIH) states that about 1% of Americans suffer from Hyperthyroidism and about 5% suffer from Hypothyroidism. From the global perspective also the classification of thyroid plays a significant role. The conditions for the diagnosis of the disease are closely linked, they have several important differences that affect diagnosis and treatment. The data for this research work is collected from the UCI repository which undergoes preprocessing. The preprocessed data is multivariate in nature. Curse of Dimensionality is followed so that the available 21 attributes is optimized to 10 attributes using Hybrid Differential Evolution Kernel Based Navie Based algorithm. The subset of data is now supplied to Kernel Based Naïve Bayes classifier algorithm in order to check for the fitness.

Keywords: *classification, curse of dimensionality, kernel based naïve bayes classifier, differential evolutionary algorithm, multivariate bayseian prediction, thyroid disease, wrapper model.*

GJCST-H Classification: *H.2.8*

Strictly as per the compliance and regulations of:

An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method

K.Geetha [α] & Capt. S. Santhosh Baboo [σ]

Abstract- Thyroid diseases are widespread worldwide. In India too, there is a significant problems caused due to thyroid diseases. Various research studies estimates that about 42 million people in India suffer from thyroid diseases [4]. There are a number of possible thyroid diseases and disorders, including thyroiditis and thyroid cancer. This paper focuses on the classification of two of the most common thyroid disorders are hyperthyroidism and hypothyroidism among the public. The National Institutes of Health (NIH) states that about 1% of Americans suffer from Hyperthyroidism and about 5% suffer from Hypothyroidism. From the global perspective also the classification of thyroid plays a significant role. The conditions for the diagnosis of the disease are closely linked, they have several important differences that affect diagnosis and treatment. The data for this research work is collected from the UCI repository which undergoes preprocessing. The preprocessed data is multivariate in nature. Curse of Dimensionality is followed so that the available 21 attributes is optimized to 10 attributes using Hybrid Differential Evolution Kernel Based Navie Based algorithm. The subset of data is now supplied to Kernel Based Naïve Bayes classifier algorithm in order to check for the fitness. This iterative process takes 21 to 25 runs until the errors are reduced or the after the errors are stabilized, the data is classified. The accuracy of classification is observed to be 97.97%.

Keywords: classification, curse of dimensionality, kernel based naïve bayes classifier, differential evolutionary algorithm, multivariate bayseian prediction, thyroid disease, wrapper model.

I. Introduction

According to a recent study published by the daily Times of India, one in ten adults in India suffers from hypothyroidism. This estimation is found on the basis of a survey conducted by Indian Thyroid Society.

The study also depicts awareness for the thyroid disease and is ranked 9th when compared to other common diseases like asthma, cholesterol, depression, diabetes, heart problem and insomania. Medical practitioners say that the symptoms of thyroid are similar to other disorders. However, the survey revealed that only 50%, of the survey population are aware of thyroid disorder, know that there are diagnostic tests for detection of this disease [3].

Thyroid disorders damage the normal functioning of the thyroid gland which causes abnormal production of hormones leading to hyperthyroidism. The occurrence of hypothyroidism in the developed world is estimated to be about 4-5%. Hypothyroidism may cause high cholesterol levels, an increase in blood pressure, cardiovascular complications, decreased fertility, and depression if not properly treated.

Hence creating awareness among the public about the symptoms and types of this disease and its diagnosis plays a crucial importance of the hour. The main objective of this research work is to show the classification of more significant features from the available raw medical dataset which helps the physician to arrive at an accurate diagnosis of Thyroid among public.

This paper is organized in such a way that section 2 elaborates about thyroid disease types, symptoms and the ill effects. Section 3 deals with the background study conducted by various authors. Section 4 focuses towards the proposed methodology of thyroid classification supported by the results and discussion in section 5.

II. Overview of Thyroid

The thyroid is an organ present in the human body and is considered to be a part of the endocrine or the hormone, system. It is located in the human neck below the Adam's apple. The main purpose of thyroid is to produce thyroid hormones. The produced hormones go through the bloodstream to all the other organs which help to control metabolism and growth development in both in adults and in children.

The thyroid looks like butterfly shape. Figure 1 shows the thyroid and its parts. The right and left lobes of the thyroid looks similar to the two wings of a butterfly. They lie on both sides of the trachea or main breathing tube. The connection between the wings is called the isthmus [5]. The thyroid gland produces hormones which primarily control human body's growth and metabolism, which means that this energy is used for all the body processes. The thyroid gland acts as an important part in breathing, blood circulation, bowel

Author α : Research Scholar Periyar University Department of Computer Science Salem, India. e-mail: kkgeetha17@gmail.com
Author σ : Associate Professor & Head, PG & Research Department of Computer Science, DG Vaishnava College, Chennai, India.
e-mail: santhos1968@gmail.com

movements, temperature of the body, muscle control, digestion, and brain function. An issue with the thyroid gland can result in problems all over the human body [6].

The thyroid gland functional data is more essential for the proper interpretation and diagnosis of the diseases associated with the gland. The principal role of the thyroid gland is to help regulation of the body's metabolism. Depending on the amount of secretion of this hormone may affect the human growth and development. When this hormone is produced very little thyroid hormone the type of disease is referred to as hypo-thyroidism. When this hormone is produced of too much it may lead to hyper-thyroidism [2].

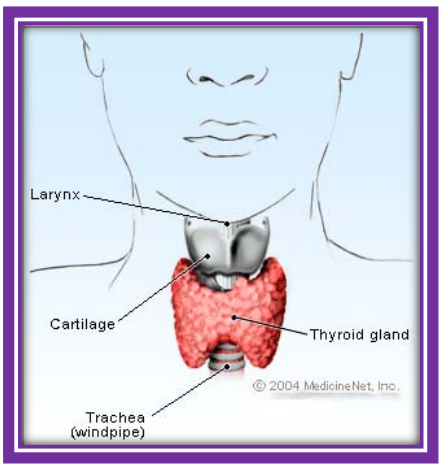

Fig. 1 : Thyroid in human body

a) Thyroid Harmones

The two hormones that are produced in the thyroid are L-thyroxine (T4) and tri-iodothyronine (T3)[5]. They regulate human body's metabolic functions such as heat generation, and the utilization of carbohydrates, proteins, and fats. Regulatory hormones from different parts of the brain control the thyroid's production of T4 and T3. In the pituitary gland, Thyrotropin-Stimulating Hormone (TSH) is released when more thyroid hormone is needed and travels via the bloodstream to the thyroid gland. TSH then stimulates the thyroid to produce T4 and T3 [5].

The pituitary gland acts like a thermostat to control the production of the hormone. When they are more in the bloodstream, the pituitary releases less TSH. When there are little in the bloodstream, the pituitary releases more TSH. With the help of this feedback system, the production of thyroid hormone is tightly controlled [5].

b) Thyroid And Health Effects

Thyroid diseases are one of the most common endocrine disorders worldwide. India too, is no exception. It is estimated that about 42 million people in India suffer from thyroid diseases [4][8].

Thyroid diseases are different from other diseases in terms of their ease of diagnosis, accessibility of medical treatment, and the relative visibility[4]. The thyroid gland secretes hormones which controls a lot of things in the human body system like metabolize the food, use energy, sleep patterns, temperature preferences, body weight balance and a lot more [7].

Both an increase and decrease in thyroid hormone production can cause health problems.

i. *Hyperthyroid*

Increase in the hormone production can cause hyperthyroidism. In medical field, “hyper” indicates too much. Hyperthyroidism crop up when the gland produces excess hormones. The most common cause for hyperthyroidism is the autoimmune disorder Graves’ disease. It is also known as an overactive thyroid, the hormone overload can cause a extensive range of physical changes. Many symptoms overlap with hypothyroidism, including thinning hair, dry skin and temperature sensitivity. The symptoms that indicate the presence of hyperthyroidism includes weight loss in spite of a good food intake, an increase in heart rate, high blood pressure, nervousness, increased sweating, enlargement in your neck, shorter menstrual periods, frequent bowel movements and trembling hands [6]. The following figure 2 shows the list of symptoms of hyperthyroid.

Fig. 2 : Symptoms of Hyperthroid

ii. *Hypothyroid*

Decrease in the hormone production can cause hypothyroidism. In medical field, .the term hypo means deficient or not enough. For example, hypoglycemia is a term for low blood sugar. Hypothyroidism is a condition that the thyroid gland does not produce required hormones. Inflammation and damage to the gland causes hypothyroidism. Weight gain or failure to lose weight despite a proper weight loss regime, lethargy,

reduced heart rate, increased cold sensitivity, numbness in hands, enlargement in the neck, dry skin and hair, heavy menstrual periods and constipation could indicate hypothyroidism. Symptoms vary from person to person, and if left untreated, they tend to worsen over time [6]. Figure 3 shows the list of symptoms of hypothyroid.

Fig. 3 : Symptoms of Hypothyroid

An increased risk of thyroid disease happens if there is a family history of thyroid disease like a type I diabetic, over 50 years of age and a stressful life [7].

Both hypothyroidism and hyperthyroidism can be diagnosed with thyroid function tests, which measures the levels of Thyroid-Stimulating Hormones (TSH) in bloodstream of human body [6].

III. Literature Review

There are many people who have studied various medical data and analyzed methods and models for preprocessing and classifying the data according to the need.

Ngan, Po Shun, et al (1999) introduced a system for discovering medical knowledge by learning Bayesian networks and rules. Evolutionary computation is used as the search algorithm. The Bayesian networks can provide an overall structure of the relationships among the attributes[13].

Ozyilmaz, Lale, and Tulay Yildirim (2002) proposed a system that includes Generalized Discriminant Analysis and Wavelet Support Vector Machine System (GDA_WSVM) method for diagnosis of thyroid diseases which includes three phases. They are feature extraction- feature reduction phase, classification phase, and test of GDA_WSVM for correct diagnosis of thyroid diseases phase, respectively [1]. The acceptable diagnosis performance of this GDA_WSVM expert system for diagnosis of thyroid diseases is estimated by using classification accuracy and confusion matrix methods, respectively. The classification accuracy of this expert system for diagnosis of thyroid diseases was obtained about 91.86% [1].

Ordonez et. al (2006) proposed a greedy algorithm to compute rule covers in order to summarize rules having the same consequent. The significance of association rules is evaluated using three metrics: support, confidence and lift [19].

Keleş, Ali, and Aytürk Keleş (2008) aims at diagnosing thyroid diseases with a expert system.In the proposed system, fuzzy rules by using neuro fuzzy method is incorporated [15].

Karaboga, D., & Basturk, B. (2008) compares the performance of ABC algorithm with that of Differential Evolution (DE), Particle Swarm Optimization (PSO) and Evolutionary Algorithm (EA) for multi-dimensional numeric problems [17].

Boryczka, Urszula (2009) focused on ant-based clustering algorithms. During the classification different metrics of dissimilarity like Euclidean, Cosine and Gower measures were used [18].

Kodaz, Halife, et al. (2009) proposed that Information gain based artificial immune recognition system (IG-AIRS) would be helpful in diagnosing thyroid function based on laboratory tests, and would open the way to various ill diagnoses support by using the recent clinical examination data. The classification used is distance-based classification systems [12].

Dogantekin et. al. (2010) introduced the diagnosis of thyroid disease. The feature reduction is performed by using Principle Component Analysis (PCA) method. The classification is done using Least Square Support Vector Machine (LS-SVM) classifier. The performance evaluation of the proposed Automatic Diagnosis System Based on Thyroid Gland ADSTG method is estimated by using classification accuracy, *k*-fold cross-validation, and confusion matrix methods respectively [14].

Karaboga et. al (2011) used ABC is used for data clustering on benchmark problems and the performance of ABC algorithm is compared with Particle Swarm Optimization (PSO) algorithm and other nine classification techniques. ABC algorithm can be efficiently used for multivariate data clustering test data sets from the UCI Machine Learning Repository are used to demonstrate the results of the techniques [10].

Stegmayer et. al (2012) proposed a novel integrated computational intelligence approach for biological data mining that involves neural networks and evolutionary computation. They used self-organizing maps for the identification of coordinated patterns variations; a new training algorithm that can include a priori biological information to obtain more biological meaningful clusters and evolutionary algorithm for the inference of unknown metabolic pathways involving the selected cluster [11].

Yeh, Wei-Chang (2012) improved simplified swarm optimization (SSO) to mine a thyroid gland dataset collected from UCI databases. Close Interval Encoding (CIE) is added to efficiently represent the rule structure, and the Orthogonal Array Test (OAT) is added to powerfully prune rules to avoid over-fitting the training dataset [16].

Chen, Hui-Ling, et al (2012) proposed expert system, Fisher Score Particle Swarm Optimization Support Vector Machines (FS-PSO-SVM) has been rigorously evaluated against the thyroid disease dataset, which is commonly used among researchers who use machine learning methods for thyroid disease diagnosis [20].

Azar et. al (2013) performed a comparison between hard and fuzzy clustering algorithms for thyroid diseases data set in order to find the optimal number of clusters. Different scalar validity measures are used in comparing the performances. K-means clustering; K-medoids clustering; Fuzzy C-means; Gustafson–Kessel algorithm; Gath–Geva algorithm clustering results for all algorithms are then visualized by the Sammon mapping method to find a low-dimensional (normally 2D or 3D) representation of a set of points distributed in a high dimensional pattern space [9].

IV. Methodology

The framework of the proposed work is shown in the figure 4. The proposed work is based on the input from the UCI repository which involves 7200 multivariate type of records. Each record has 21 attributes. Out of the 21 attributes 15 are continuous data and 6 are discrete data.

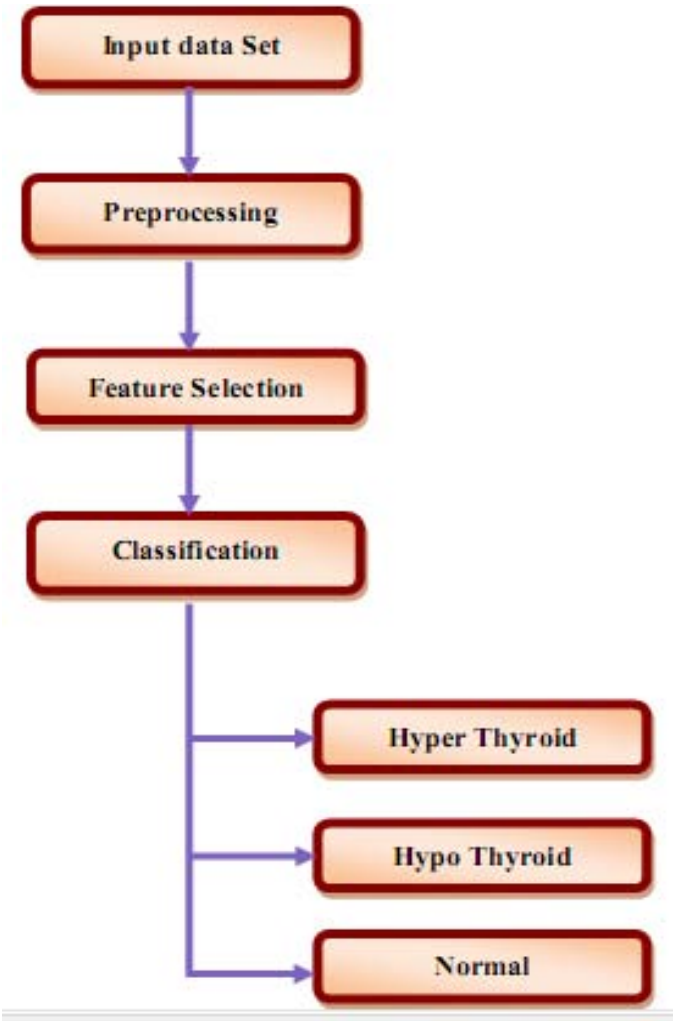

Fig. 4 : Framework of the proposed method

The following steps are involved in the process of the proposed work.

1. The data taken from UCI repository undergoes preprocessing where missing value and not a number constraint are checked using masking method. If the missing value or Not a Number (NaN) values are present it is replaced by the mean value of the column.
2. The preprocessed data is fed into a hybrid algorithm termed as Differential Evolution (DE). This algorithm

is used for creating subset of child from the parent records.

3. The subsets of data are applied to Kernel Based Bayesian classifier algorithm to check the fitness. The fitness is measured by error stabilization.
4. After stabilization is achieved , the data is classified into 3 classes as

a. Hypo Thyroid
b. Hyper Thyroid
c. Normal

a) Data Set

The following table 1 shows the characteristics of Data set collected from UCI repository.

Table 1: Characteristics of Dataset

Data Set Characteristics	Multivariate, Domain-Theory
Attribute Characteristics	Categorical, Real
Associated Tasks	Classification
Number of Instances	7200
Number of Attributes	21 (Continuous -15; Discrete -6)

To ensure that the patterns derived are as accurate as possible, it is essential to improve the quality of the datasets in the pre-processing stage. Most real life data sets contain a certain amount of redundant data, which does not contribute significantly to the formation of important relationships. This redundancy not only increases the dimensionality of the data set and slows down the data mining process but also affects the subsequent classification performance [21].

Attribute selection is the process of removing the redundant attributes that are deemed irrelevant to the data mining task. However, the presence of attributes that are not useful to classification might interfere with the relevant attributes to degrade classification performance. This is due to the noise that is contributed by these additional attributes and raises the level of difficult [21].

The objective of attribute selection is therefore to search for a worthy set of attributes that produce comparable classification results to the case when all the attributes are used. In addition, a smaller set of attributes also creates less complicated patterns, which are easily comprehensible, and even visualized, by humans [21].

It has to be noted that for a data set with n attributes, there are 2^n–1 possible subsets. Therefore, an exhaustive search for an optimal set of attributes would be time-consuming and computationally expensive if n is large [21].

b) Preprocessing

The pre-processing step is necessary to resolve several types of problems including noisy data, redundant data, missing data values, etc. The high quality data will lead to high quality results and reduced costs for data mining. Missing data should be pre-processed so as to allow the whole data set to be processed by a required algorithm. Moreover, most of the existing algorithms are able to extract knowledge from data set that store discrete features. If the features are continuous, the algorithms can be integrated to create discrete attributes [22].

In the proposed work, the data taken from UCI repository has both continuous and discrete data which undergoes preprocessing. In this stage, the missing value and not a number constraint are checked using masking method. If the missing value or Not a Number (NaN) values are present it is replaced by the mean value of the column.

c) Feature Selection

Accurate diagnosis of diseases and subsequently, providing efficient treatment, forms an important part of valuable medical services given for patients in the health-care system. The unique characteristics of medical databases that pose challenges for data mining are the privacy-sensitive, heterogeneous, and voluminous data. These data may have valuable information which awaits extraction. The required knowledge is found to be encapsulated in/as various regularities and patterns that may not be evident in the raw data or the preprocessed data.

Extracting knowledge has proved to be priceless for future medical decision making. Feature selection is crucial for analyzing various dimensional bio-medical data. It is difficult for the biologists or doctors to examine the whole feature-space obtained through clinical laboratories at one time. All the computational algorithms recommend only few significant features for disease diagnosis. Then these recommended significant features may help doctors or experts to understand the biomedical mechanism better with a deeper knowledge about the cause of disease and provide the fastest diagnosis for recovering the infected patients as early as possible [24].

Feature selection methods tend to identify the features most relevant for classification and can be broadly categorized as either subset selection methods or ranking methods. The former type returns a subset of the original set of features which are considered to be the most important for classification [24][26].

Feature selection, is an effective in dimensionality reduction, by removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility [24][25]. Feature selection algorithms generally fall into two broad categories. They are:

A. The filter model
B. The wrapper model

i. *Filter Model*

The filter model depends on general characteristics of the training data to select some features without involving any learning algorithm. The

filter model assesses the relevance of features from data alone, independent of classifiers, using measures like distance, information, dependency (correlation), and consistency [24][25].

ii. *Wrapper Model*

The wrapper model needs one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. For each of the generated new subset of features, the wrapper model is supposed to learn the hypothesis of a classifier. It has a propensity to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it also tends to take more computation time and is more expensive than the filter model [24][25].

This research work uses the Wrapper model for feature selection. In wrapper methods, the algorithm that selects the features uses a classification algorithm for evaluation. Accordingly, wrapper methods are more precise but computationally more complex [27][28], and they also depend on the data selected for classifier development. Since these data guide the selection, they can lead to over-fitting [28][29]. A broad spectrum of various wrappers is used in today's approaches. For example, the forward and the backward floating search and their combinations are commonly used, where one feature is added or reduced at a time, depending on the classification accuracy, Evolutionary algorithms are also used.

iii. *Defferential Evolution*

Differential Evolution, or briefly DE [28][30][31][32] is a simple but effective search method for continuous optimization problems. According to Xinjie and Mitsuo (2010), DE represents a direction based search that maintains a vector population of candidate solutions. Like other usual Evolutionary Algorithms (EAs), it uses mutation, crossover and selection. The key part of DE, which differentiates it from standard EAs, is the mutation operator that perturbs the selected vector according to the scaled difference of the other two members of the population. The operation of DE is shown as pseudo-code in Algorithm 1.

Algorithm 1: Differential Evolution (DE)- pseudo-code.

1: Initialization and parameter setting
2: while termination condition not met do
3: for all population member—vector vi do
4: create mutant vector u_i
5: crossover vi and u_i to create trial vector t_i
6: end for
7: for all population member—vector v_i do
8: if $f(t_i) \leq f(v_i)$ then
9: $v_i \leftarrow t_i$
10: end if

The population of size NP contains vectors and each vector v_i, of dimensionality D, consists of real-valued parameters, $v_i = (v^1_i, \ldots, v^D_i) \in RD$, 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} represent the lower and upper bound, respectively. In each generation, a new population is created through mutation and crossover. This new population is composed of the trial vectors t_i. For each member of the current population, v_i (called the target vector), a new corresponding mutant vector u_i is formed using mutation. The mutation is conducted according to

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

Here u_i is a mutant while vr1, vr2 and vr3 are population vectors selected randomly with the condition $i \neq r1 \neq r2 \neq r3$, and $F \in [0,\infty)$ is the scale factor which represents a parameter of the algorithm. After the mutation, crossover occurs between the target vector v_i and the corresponding mutant u_i creating a trial vector t_i. The crossover is done as follows:

$$tji = \begin{cases} u^j_i & \text{if } U[0, 1) \leq CR \text{ or } j = r_j, \\ v^j_i & \text{otherwise} \end{cases}$$

for j = 1, . . .,D. Here t_i is a trial vector obtained through crossover, U[0, 1) is a variable with its value randomly selected from the interval [0, 1) with uniform distribution, r_j is a random variable with the value from the set {1, . . .,D}, while $CR \in [0, 1)$ is the crossover rate and represents a parameter of the algorithm. The described crossover is called the binomial crossover. Once the trial vector population has been created, vectors that transfer over to the next generation, i.e., which will constitute the new population, are selected. A given trial vector t_i replaces the corresponding target vector v_i if it is of equal or lesser cost, according to the given objective/fitness function. Due to its simplicity, DE is a very popular search method that has been successfully applied to various problems [28].

d) Classification

A classifier is a function f that maps input feature vectors $x \in X$ to output class labels $y \in \{1, \ldots, C\}$, where X is the feature space [33].

i. *Kernel Based Naïve Bayian Classifier*

Naive Bayesian Classifier is a simple probabilistic classifier with an assumption of conditional independence among the features, i.e., the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. It only requires a small amount of training data to estimate the parameters necessary for classification. Many experiments have demonstrated that NB classifier has worked quite well in various complex real-world situations and outperforms many other classifiers. Kernel estimation has been used in cases of datasets with numerical attributes [23][24].

The Naive Bayes classifier classifies data in two steps:

1. Training step: Using the training data, the method estimates the parameters of a probability distribution, with the assumption that the predictors are conditionally independent given the class.
2. Prediction step: For any unknown test data, the method computes the posterior probability of the sample belonging to each class. The method then classifies the test data.

The class-conditional independence assumption simplifies the training step and the estimate of the one-dimensional class-conditional density for each class is predicted individually [35].

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods [34].

Naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector

$$\mathbf{x} = (x_1, \ldots, x_n)$$

representing some *n* features (independent variables), it assigns to this instance probabilities

$$p(C_k|x_1, \ldots, x_n)$$

for each of *K* possible classes [36][37]. If the number of features *n* is large, the conditional probability can be decomposed as

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

The advantage of using Naive Bayes algorithm are

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction [34].
- When assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression and less training data is required[34].

V. Results and Discussion

The proposed model is developed using Matlab. The proposed work is designed to have two panels. One is the display panel and the other is analysis panel. The following figure 5 shows the framework of the Evolutionary Multivariate Bayesian prediction method where the data is loaded from the repository. Figure 6 and figure 7 displays the preprocessing and feature selection stage respectively.

Data distribution and error stabilization is shown in display panel of figure 8. The classification is evaluated based on ten evaluation metrics whose values are shown in the figure 9. 21 epohs (runs) are carried out for the data. After stabilization is achieved , the data is classified into 3 classes.

1. Hypo Thyroid
2. Hyper Thyroid
3. Normal

The accuracy of classification is achieved as 97.97%.

With the aim of accessing the classifier and to compare the output classes, Receiver Operating Characteristic (ROC) is used. The ROC is shown in the figure 10 which compare each class of thyroid based on their True Positive rate and False Positive rate. The number of instances of Hyper thyroid and hypothyroid among male and female is shown in figure 11.

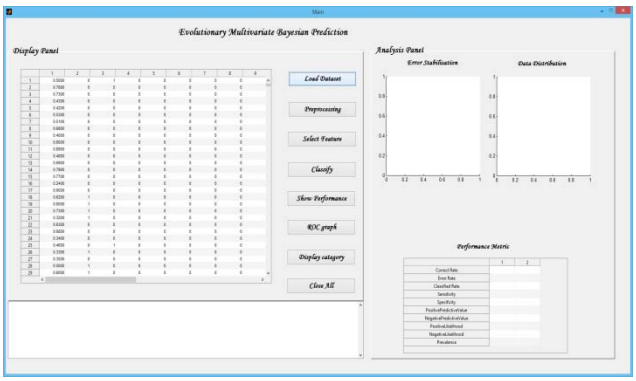

Fig. 5 : Data Load from the repository

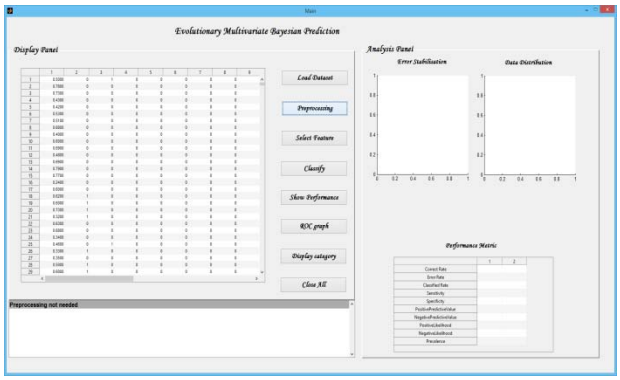

Fig. 6 : Preprocessing

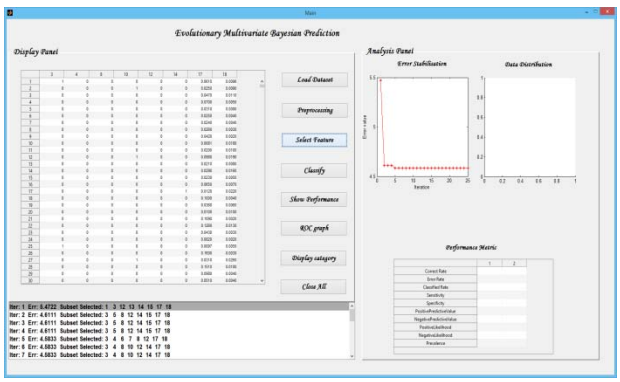

Fig. 7 : Feature Selection

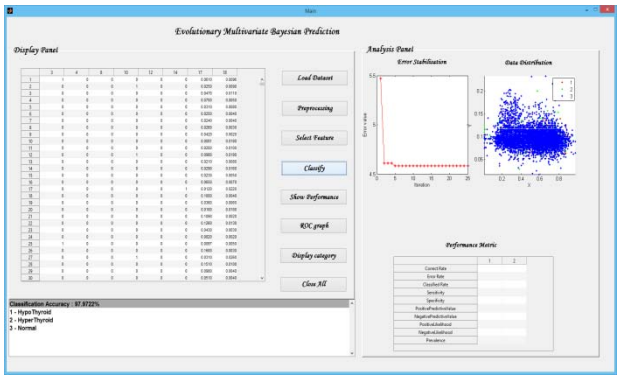

Fig. 8 : Classification

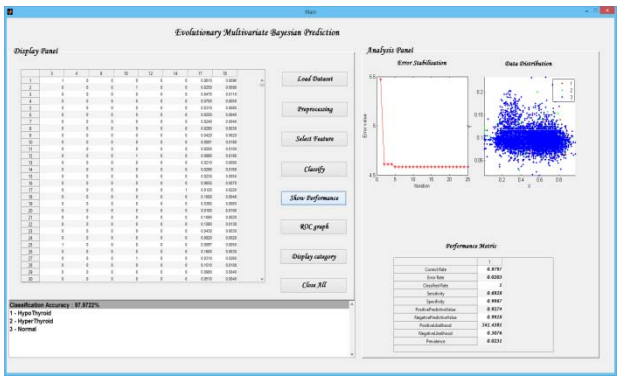

Fig. 9 : Performance metrics

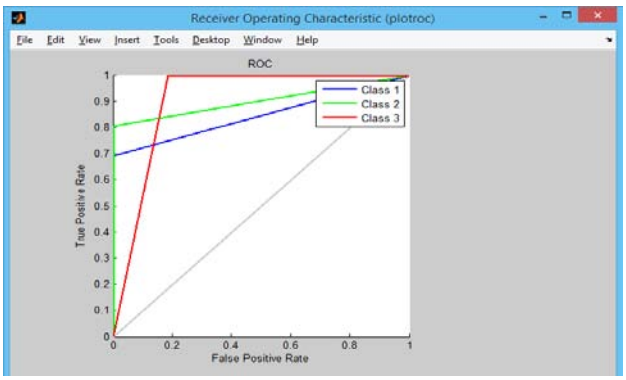

Fig. 10 : Receiver Operating Characteristic (ROC)

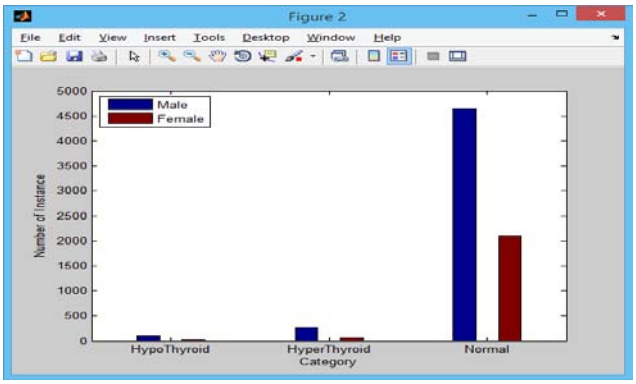

Fig. 11 : Number of instances of Hyper thyroid, Hypo thyroid among male and female

VI. Conclusion

The objective of this research work is aimed to show the classes of thyroid from the available raw medical dataset helps the physician to arrive at an accurate diagnosis. The results show that the proposed Evolutionary Multivariate Bayisean Prediction classifier model achieves remarkable dimensionality reduction from among the 7200 medical datasets obtained from the UCI repository with 21 attributes (Continuous -15; Discrete - 6). 21 epohs (runs) are carried out for the data and after stabilization, the data are classified as Hyper, Hypo and Normal classes. The results are evaluated based on ten evaluation metrics and the accuracy of classification is 97.97%.

References Références Referencias

1. Ozyilmaz, Lale, and Tulay Yildirim. "Diagnosis of thyroid disease using artificial neural network methods." *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*. Vol. 4. IEEE, 2002.
2. Ozyilmaz, Lale, and Tulay Yildirim. "Diagnosis of thyroid disease using artificial neural network methods." Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on. Vol. 4. IEEE, 2002.
3. http://timesofindia.indiatimes.com/india/1-in-10 - Indians-suffer-from-thyroid-disorder-Studyarticlesho w / 46007453.cms [Accessed Dec 2015]
4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169866/(accessed dec 2015)
5. http://www.emedicinehealth.com/thyroid_faqs/article_em.htm
6. http://www.foxnews.com/health/2012/02/10/ hypo thyroidism -versus-hyperthyroidism.html (accessed dec 2015)
7. http://www.thehealthsite.com/diseasesconditions / world- thyroid – day – 2012 - facts-you-should-know/(accessed dec 2015)
8. http://www.ias.ac.in/currsci/oct252000/n% 20 kochu pillai.PDF (Accessed dec 2015)
9. Azar, Ahmad Taher, Shaimaa Ahmed El-Said, and Aboul Ella Hassanien. "Fuzzy and hard clustering analysis for thyroid disease." Computer methods and programs in biomedicine 111.1 (2013): 1-16.
10. Karaboga, Dervis, and Celal Ozturk. "A novel clustering approach: Artificial Bee Colony (ABC) algorithm." Applied soft computing 11.1 (2011): 652-657.
11. Stegmayer, Georgina, Matias Gerard, and Diego H. Milone. "Data mining over biological datasets: An integrated approach based on computational intelligence." Computational Intelligence Magazine, IEEE 7.4 (2012): 22-34.
12. Kodaz, H., Özşen, S., Arslan, A., & Güneş, S. (2009). Medical application of information gain

based artificial immune recognition system (AIRS): Diagnosis of thyroid disease. Expert Systems with Applications, 36(2), 3086-3092.
13. Ngan, P. S., Wong, M. L., Lam, W., Leung, K. S., & Cheng, J. C. (1999). Medical data mining using evolutionary computation. Artificial Intelligence in Medicine, 16(1), 73-96.
14. Dogantekin, Esin, Akif Dogantekin, and Derya Avci. "An automatic diagnosis system based on thyroid gland: ADSTG." Expert Systems with Applications 37.9 (2010): 6368-6372.
15. Keleş, Ali, and Aytürk Keleş. "ESTDD: Expert system for thyroid diseases diagnosis." Expert Systems with Applications 34.1 (2008): 242-246.
16. Yeh, Wei-Chang. "Novel swarm optimization for mining classification rules on thyroid gland data." Information Sciences 197 (2012): 65-76.
17. Karaboga, Dervis, and Bahriye Basturk. "On the performance of artificial bee colony (ABC) algorithm." Applied soft computing 8.1 (2008): 687-697.
18. Boryczka, Urszula. "Finding groups in data: Cluster analysis with ants." Applied Soft Computing 9.1 (2009): 61-70.
19. Ordonez, Carlos, Norberto Ezquerra, and Cesar A. Santana. "Constraining and summarizing association rules in medical data." Knowledge and Information Systems 9.3 (2006): 1-2.
20. Chen, H. L., Yang, B., Wang, G., Liu, J., Chen, Y. D., & Liu, D. Y. (2012). A three-stage expert system based on support vector machines for thyroid disease diagnosis. Journal of medical systems, 36(3), 1953-1963.
21. Tan, Kay Chen, et al. "A hybrid evolutionary algorithm for attribute selection in data mining." Expert Systems with Applications 36.4 (2009): 8616-8630.
22. Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas. "Data preprocessing for supervised leaning." International Journal of Computer Science 1.2 (2006): 111-117.
23. Witten, H.I., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA, USA.
24. Sasikala, S., S. Appavu alias Balamurugan, and S. Geetha. "Multi filtration feature selection (MFFS) to improve discriminatory ability in clinical data set." *Applied Computing and Informatics* (2014).
25. Xing, E., Jordan, M., Karp, R., 2001. Feature selection for high-dimensional genomic microarray data. In:Brodely, C.E., Danyluk, A.P. (Eds.), Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco,CA, USA, pp. 601–608.
26. Jazzar, M.M., Muhammad, G., 2013. Feature selection based verification/identification system using fingerprints and palm print. Arabian J. Sci. Eng. 38 (4), 849–857.
27. Wang, G., Jian, M. and Yang, S. (2011). IGF-bagging: Information gain based feature selection for bagging, International Journal of Innovative Computing, Information and Control 7(11): 6247–6259.
28. Martinoyić, Goran, Drażen Bajer, and Bruno Zorić. "A differential evolution approach to dimensionality reduction for classification needs." *International Journal of Applied Mathematics and Computer Science* 24.1 (2014): 111-122.
29. Loughrey, J. and Cunningham, P. (2004). Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets, in M. Bramer, F. Coenen and T. Allen (Eds.), The Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, Berlin/Heidelberg, pp. 33–43.
30. Storn, R. and Price, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization 11(4): 341–359.
31. Price, K.V., Storn, R.M. and Lampinen, J.A. (2005). Differential Evolution. A Practical Approach to Global Optimization, Springer-Verlag, Berlin/Heidelberg.
32. Xinjie, Y. and Mitsuo, G. (2010). Introduction to Evolutionary Algorithms, Springer-Verlag, London.
33. Murphy, Kevin P. "Naive bayes classifiers." *University of British Columbia* (2006).
34. http://www.analyticsvidhya.com/blog/2015/09 /naive -bayes-explained/ (Accessed dec 2015)
35. http://in.mathworks.com/help/stats/naive-bayes-classification.html (Accessed dec 2015)
36. https://en.wikipedia.org/wiki/Naive_Bayes _classifier (Accessed dec 2015)
37. *Narasimha Murty, M.; Susheela Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.* ISBN 0857294946.

This page is intentionally left blank