Chapter 1

Introduction

Contents

1.1 Introduction

An interesting and wide range of applied mathematics research is being done in the field of medical sciences. Researchers of applied mathematics use soft computing techniques as a tool in solving the problems of medical sciences. The Thesis is related to diagnose some regular (common) dermatological disorders using various soft computing techniques specially kernel based techniques.

People in India, particularly in rural area and small town, are not much conscious about skin diseases and do not consult dermatologist at the initial stage. Unfortunately, in most of the rural areas dermatologists are hardly available and people are generally treated by general practitioners or paramedical staff at primary health centres, community health centres and referral hospitals. Many dermatological diseases such as Bacterial Infections, Fungal Infections, Eczema and Scabies have very similar features. So, to diagnose these diseases at the initial stage by non derma-

tologist is very difficult. Because of the improper diagnosis many times patients are treated incorrectly by the mixture of antibacterial, antifungal and steroid preparation locally. Such treatment is perilous to the patient and it precipitates chances for relapse and side effects of local agents like steroids. There is definite need for proper diagnosis and treatment for such disorders. Modern medicine is looking for solutions, which help the doctors taking decisions. Machine learning techniques provide great support to the doctors to diagnose diseases. It will make doctors more confident about their decision.

Related Work in Skin Disorders: More than a thousand diseases are observed in dermatology and they are classified in different types of disorders. Lots of research work has been carried out in diagnosis of various skin diseases. Bapko *et al.* have developed diagnostic system with 90% success rate using Artificial Neural Network to diagnose 17 types of skin diseases such as scabies, Acne, Vulgari etc [7]. Naser & Akkila have developed an expert system using CLIPS (C Language Integrated Production System) to diagnose 9 different skin diseases namely, Psoriasis, Eczema, Ichthyosis etc [90]. Major work has been done in diagnosis of differentiate Erythemato-Squamous disease and Skin cancer. Guvenir & Emeksiz have used three algorithms namely, Nearest Neighbor Classifier (NN), Naive Bayesian Classifier (NBC) and Voting Feature Intervals-5 (VFI5) to diagnose Erythemato-Squamous diseases [44]. For effective implementation of NN algorithm they gave weights to each feature using Genetic Algoritm(GA). Karlik & Harman have used Artificial Neural Network using Back Propagation algorithm for early diagnosis of Erythemato-Squamous diseases and achieved 98 % success for 6 Erythemato-Squamous diseases [66]. Danjuma & Osofisan have used the same dataset and diagnose the Erythemato-Squamous diseases using Naive Bayes algorithm, Multilayer Perceptron and J48 Decision tree and got accuracy of 97.4 %, 96.6 % and 93.5% respectively [28]. Olatunji & Hossain diagnose the Erythemato-Squamous diseases using Extreme Learning Machine(ELM) [94]. Many researchers have classified the diseases after applying various feature selection techniques ([5], [39], [101], [136], [137]). Lau *et.al.* have used 3 layers back-propagation Neural Network classifier for automatic early detection of Skin cancer [74]. Esteva *et. al.* have diagnose Skin Cancers using deep convolution neural network [35]. Bakheet have developed computer aided diagnostic system using Support Vector Machine (SVM) to optimize set of HOG (Histogram of Oriented Gradient) based descriptors of skin lesions [6].

As mentioned in the above literature survey, most of the researchers have focused on the diagnosis of Erythemato-Squamous disease and Skin Cancer. But, very less emphasis was given on diagnosis of very common skin diseases such as Bacterial Infections, Fungal Infections, Eczema and Scabies. Our purpose is to develop some techniques which help general medical professional for proper diagnosis of these diseases and serve as a second opinion for dermatologist. For this purpose dataset was collected as mentioned below.

1.2 Data Acquisition

Dataset was collected for Bacterial skin infections, Fungal skin infections, Eczema and Scabies from The H M Patel Centre for Medical Care and Education (Shree Krishna Hospital), Karamsad, Gujarat, India. Firstly, a research proposal (according to the Human Research Ethics Committee (HREC) format of The H M Patel Centre for Medical Care and Education) was prepared to get the permission to collect data from their Skin and V.D. Department. The research proposal is presented in the HREC Committee meeting. They asked for some modifications to the research proposal in accordance with the conditions of HREC. After incorporating all the required changes, the research proposal got accepted in the second meeting. Then, a detailed proforma is prepared for the data collection of the above mentioned skin diseases under the guidance of Head of the Department of Skin and V. D., Dr. Rita Vora. The data were collected during OPD through one-on-one interview with the patients. Patients consent was taken before beginning the interview. Out of the 47 features in the proforma, 11 features were non-clinical features and the rest were clinical features. The information regarding non-clinical features were filled up by me after interviewing the patients and the same for clinical features were done by the dermatologists of the department namely, Dr. Rita Vora and Dr. RahulKrishna. Data was collected for a total of 470 patients. Out of these, 139 patients had Bacterial skin infections, 146 had Fungal skin infections, 98 had Eeczema and 87 had Scabies. The answers to the questions in the proforma were in the form of Yes or No i.e. whether the feature was present or not. The soft copy of each of the 470 forms was prepared in which the features were expressed in binary form. 1 indicated that feature is present and 0 indicated that feature is absent. The diseases were expressed in numerical form in the soft copy, where 1 indicated Bacterial infection, 2 indicated

Fungal infection, 3 indicated Eczema and 4 indicated Scabies. After preparing the soft copy of these 470 forms, random partitions of 80%-20%, 70%-30% and 60%-40% for training and testing purpose were created. Approval Letter, Proforma and Certificate of data collection are attached in the Appendix-A for reference. Detailed description of the Datasets are given in the Appendix-A.

1.3 Methodologies for Classification

The problem is to diagnose the above mentioned skin diseases but mathematically it is a classification problem. For this purpose we have used various soft computing techniques. Components of soft computing includes machine learning techniques such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Extreme Learning Machine (ELM) etc. Soft computing techniques simulate biological processes. Basic feature of soft computing techniques is to train the network which gives minimum error between actual output and network output. So, fundamentally all soft computing techniques have strong bonding with mathematical optimization.

The following methodologies (soft computing techniques)have been used in classification of the above mentioned skin diseases.

(I) Artificial Neural Network for Classification

An Artifical Neural Network(ANN) is a non-parametric, non linear computational model. It has provided the first step towards the Artificial Intelligence. It simulates the working of brain. When data represent non linear relationship, ANN methodologies are very much useful compared to classical statistical techniques. Due to its effectiveness in pattern recognition, it is widely used in medical field. The history of ANN started in 1890 with the paper of William James about scientific attempt to study brain activity patterns [65]. In 1943 McCulloch and Pitts had created simple model of neural network to describe how neurons in brain might work [86]. But, ANN became popular from 1986, when multiple layered neural networks with Back Propagation learning rule was introduced by Rumelhart *et al.* to separate non linear data [113].

In Multilayer Perceptron there is one input layer, one output layer and one or more hidden layers. During training of the network, training data are applied to the input layer and initial weights are provided. A forward sweep is made through the network and output of each element is computed layer by layer. The difference between the calculated output of the final layer (output layer) and the desired output is calculated which is called error. If this error is greater than tolerance then weights are updated using some optimization techniques such as Steepest Descent method, Levenberg Marquardt optimization technique etc. Weights from output layer to input layer are updated by back propagation algorithm.

To classify the above mentioned skin diseases we used ANN with back propagation learning rule. To minimize the network error, instead of most popular steepest descent method, we have used the Levenberg-Marquardt optimization technique, which has more potential to converge. The classification accuracy is measured using various accuracy measures taking one and two hidden layers. Using two hidden layered network a very good classification accuracies of 96.23% and 97.17% for 70%-30% and 80%-20% training - testing data partitions respectively are achieved by normal accuracy measure. Also, for the same data partitioned 94.23% and 95.70% accuracies are achieved using F-score accuracy measure.

(II) Support Vector Machine with Positive Definite Kernels

Support Vector Machine(SVM) is the most widely used soft computing technique for non linear data. It's good generalization ability, better performance and robust mathematical theory makes it popular among other machine learning algorithm. It is the combination of machine learning theory, optimization algorithms from operation research and kernel functions from functional analysis. It is also known as large margin classifier.

The history of SVM started in the beginning of 1960s when Vapnik *et. al.* have developed an algorithm to construct an optimal separating hyperplane for separable data [127]. In 1992 Boser *et. al.* have constructed the optimal separating hyperplane in Hilbert space using Mercer's theorem, which explicitly map the input vectors into higher dimension Hilbert space [12]. In 1995 Cortes *et. al.* generalized the maximum margin classifier in feature space for nonseparable data, which is known as Support Vector Machine(SVM) [22].

SVM is used to diagnose some regular skin diseases. From the Dataset-I (Appendix-A) 70% of data is randomly selected as training data while rest of 30% data is used for testing. SVM is basically developed for binary classification problem. Here it is used for four classes. Also, since kernel plays a very important role in the performance of SVM, classification is carried out with various positive definite kernels. We use LIBSVM 3.20 [76]. Linear Kernel, Polynomial Kernel and Radial Basis Kernel Function (RBF) are in built Kernel functions in LIBSVM. Apart from this we have also used t-student and Inverse Multiqadratic Kernel function. For each kernel, parameters are set using grid search method (2.2.2) and validation is done using 10 fold validation method (2.4.8). Results exhibits that 95.39%, 90.78% and 93.80% classification accuracies are achieved using normal accuracy measures, F-score measure and G-score measure respectively using RBF Kernel and Polynomial Kernel for Dataset-I.

(III) Support Vector Machine with Novel Indefinite Kernel

For classification of non linear data, Support Vector Machine (SVM) is better classifier with appropriate choice of kernel. Measuring similarities among training samples is the fundamental step of classification. Kernels measure similarity between samples and provide classification of data. Though positive definiteness is the traditional requirement of SVM, in many applications indefinite kernels are known to yield better classification accuracy.

Unlike Mercer's kernels, indefinite kernels are defined in an inner product space endowed with a Hilbert topology called Kreĭn space, which is a pseudo Euclidian space [95]. Since, the kernels are indefinite, instead of minimizing the error, the emphasize is on stabilizing process.

This study develops a novel indefinite kernel which is a Modified Gaussian Kernel. The kernel is indefinite defined in pseudo Euclidean space. This kernel is used in diagnosis of some regular skin disorders comprising in Dataset-I and Erythemato-Squamous Disease (ESD) discussed in Dataset-II. The classification accuracy achieved by the novel kernel function is 91.50% and 98.63% for Dataset-I and Dataset-II respectively, which is better than that obtained by traditional Mercer's kernels [102]. Similar analysis has been carried out by considering various types of non Euclidean

distances in Radial Basis Kernel Function and non standard inner products in Polynomial kernel. Positive definite/indefinite property of the various kernels are also analyzed. The results exhibits good classification accuracies by these kernels.

(IV) Probabilistic Approach in Feature Selection

In many datasets, number of features are very high compare to the number of available samples for training. This increase the dimensionality of the dataset. If we use such dataset for classification, the classifier may face the problem of over fitting and good generalisation may not be obtained. So, it is necessary to remove these types of features from the dataset before applying any classification technique. It will reduce the dimensionality of the dataset and hence help in achieving good generalized classification accuracy. Statistical analysis reveal high correlation among some features which can be exploited to reduce the dimensionality of the dataset. There are three methods for feature selection: Filter method, Wrapper method and Embedded method.

In this study dermatological disorders discussed in Dataset-I and Dataset-II are diagnosed using a novel approach of feature selection. To overcome drawbacks of Filter Techniques and Wrapper Techniques we have combined both techniques which we call Filter technique and Partial Forward Search (F_PFS) algorithm. The method works for both balanced and imbalanced datasets as well as binary and multiclass datasets. It also works either inputs are binary or it is given in some scale according to the intensity of a feature.

A comparison analysis of F_PFS algorithm with IFSFS method [136] has been carried out. The results show that the new approach of feature selection i.e. F_PFS algorithm have reduced 26% features from Dataset-I and 39% features from Dataset-II with good classification accuracies of 89.36% and 97.27% for 70%-30% data partitions of the respective Datasets. The accuracy obtained by IFSFS method on Dataset-II for 70%-30% training-testing data partitions is 94.44%. This reveal the effectiveness of F_PFS algorithm. For comparison purpose IFSFS method is applied on Dataset-I. But, same accuracy is achieved for the model which contains more number of features than the model selected by F_PFS algorithm. The study deduced that using F_PFS algorithm better classification accuracy is achieved with the reduced dimensionality of the model and comparatively less effort is required due to threshold technique used in the algorithm.

(V) Kernel Based Extreme Learning Machine for Classification

The major pitfall of feedforward neural networks is their slow speed due to gradient descent based learning methods. In 2004 Haung *et.al.* have proposed a new learning algorithm called Extreme Learning Machine(ELM) which is a single layer feedforward neural network [53]. ELM is based on empirical risk minimization theory [32]. The major advantage of ELM is, the weights need not to be trained hence learning is very fast.

In this method, the rectangular system of linear equations to obtain network weights can be solved using the Moore-Penrose inverse. Therefore there is no requirement to train the network to find weights, instead they can be obtained analytically. This enables the algorithm very fast. The Moore-Penrose inverse gives minimum norm least-squares solution of a linear system [121]. Hence, the network weights are not only minimum but they are smallest norm that gives unique solution. So, the best generalised solution is obtained.

Most of the researchers are using popular Radial Basis Function and Polynomial Kernel Function in ELM. But, in this study Chi-Square kernel is used in ELM and better classification accuracy than the above mentioned kernels are achieved for Dataset-I with very less computational time ([100]). Comparative study of Extreme Learning Machine with Artificial Neural Network and Support Vector Machine is also carried out. It has been observed that good generalized performance (92.91% classification accuracy) with extremely high speed (0.0324 seconds) is achieved using Exponential Chi-Square kernel in ELM to diagnose skin diseases.

1.4 Layout of the Thesis

The Thesis is organized as follows:

Chapter 1 deals with general introduction of the thesis.

Chapter 2 focuses on the Mathematical Preliminaries.

In Chapter 3 we have used Back Propagation algorithm to update weights of the network, in which Levenberg-Marquardt algorithm is used to minimize network error. Different hidden layers in the network are used for classification of four common skin diseases. The results exhibit a very good classification accuracy 97.17% for Dataset-I.

Chapter 4, discusses the Support Vector Machine with various Mercer's Kernels to diagnose several skin diseases and 95.39% classification accuracy is achieved. The classification accuracy is also computed by different statistical measures and different training-testing data partitions of Dataset-I using different kernel functions.

Indefinite kernel in Kreĭn Space is being studied in Chapter 5. A novel indefinite kernel which gives better classification accuracy for Dataset-I under study is proposed. Various distances and different inner products in Radial Basis Kernel function and Polynomial Kernel function are used to diagnose skin disorders. Positive definite/Indefinite property of these similarity measures are determined using Gram matrix.

A new algorithm of feature selection, named Filter Technique and Partial Forward Search (F_PFS) algorithm has been proposed in Chapter 6. This algorithm is applied on two different skin datasets, Dataset-I & Dataset-II. The algorithm provides good classification accuracies on both datasets with reduced set of features.

Chapter 7, focuses on the Extreme Learning Machine(ELM) with different kernels for classification. It gives global optimum with extremely high speed. Exponential Chi-Square kernel gives good classification accuracy. A study of comparison of accuracy and learning speed of Artificial Neural Network, Support Vector Machine and Extreme Learning Machine has been carried out.