CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

Extracting meaningful information from tons of data is an art. Volume of data is increasing in every domain of life. Especially in the medical domain, data is created in the form of patient history, treatment records, post-treatment records, billing records, history records so on and so forth. Transforming this data into useful information to the medical practitioner is an important facet of intelligence. Processing this vast amount of data is beyond human capacity.

Data mining is a many faceted joint effort from databases, machine learning, statistics, artificial intelligence, text mining. This joint effort has enabled in molding mountains of data into nuggets of gold, i.e., valuable information. Machine learning sometimes conflated with data mining, is a scientific discipline. This deal with building the algorithms and studying that learns from the data by building an appropriate model. Arthur Samuel defined machine learning as *"Field of study that gives computers the ability to learn without being explicitly programmed"*. Machines are embedded with artificial knowledge to think as human beings think. In a way machine learning uses data mining and data mining uses machine learning for a better performance.

In the recent past lot of research is directed towards the medical field as it plays a vital role in every facet of human life. Pollution, natural calamities, industrial wastage, changing weather conditions etc., lead to severe health calamities. Human kind is affected from unexpected virus, severe health conditions. Cancer is one major virus spreading among all walks of human life. Even children are not spared with this disease. Medical practitioners have a tough task in identifying the risks, causes, symptoms of the disease. There is no common treatment modality for these diseases.

Research groups are focusing on this dimension. The most common form of cancer among women is cervical cancer in developing countries. It ranks the 2nd most common form of cancer among the rural women in India. Women play multiple roles in the society and family at large. The development of the society and the family depends on the abilities, capabilities and most of all the physical fitness of women. A weak woman will produce weak generation. The society needs to build a strong and a healthy nation. The back bone behind this is the woman. Women need to keep an eye on their physical fitness; hence a regular check up is essential. If this virus is not identified at a very early stage it will bring chaos into the families. “Only I can change my life. No one can do it for me” said Carol Burnett, Women’s Health, and January 2014.

Cervical cancer has been picked as the main domain of study in this research as it stands second most prevalent disease among women in India and also in the developing countries as per Information Centre for HPV and Cancer (ICO), 2014. The rate of diagnosis of the cancer is very low as proper technology is not being used by the medical practitioners in identifying the most influential risk factors for the spread of the disease. Data mining would be a helpful tool in analyzing the vast amount of data and extracting useful patterns as an aid in identifying the disease.

Data mining is often used for finding the hidden information from a database. Data mining comprises of different algorithms in order to accomplish varied kinds of tasks. The basic motive of these algorithms is to fit a model to the data, determine a model which is very close based on the characteristics or features of the data. Three parts of data mining algorithms would be: *model, preference and search.* Fitting the model to the data is the major purpose of any algorithm. In order to fit a model over the other data some criteria would be used as *preferences*. All the algorithms use some kind of *search* technique to extract the data. The designed model can be either a predictive kind or descriptive kind. Predictive modeling may be made based on the use of historical data, like predicting the weather forecast. A descriptive model gives description by identifying the pattern in the data or compare relationships in the data.

Data mining techniques when applied on medical data takes the past data for analysis and generates a predictive model that helps in identifying the trends of the disease [103]. The medical data comes in different forms as different practitioners use different ways of storing the data. This data need to be cleansed, processed in order to be analyzed using the data mining algorithms.

Data cleansing and transformation play a very major role for accurate analysis of the data, specifically medical data. Once the data is ready for analysis, different techniques could be applied based on the type of analysis and the outcome purposed. The various forms of functionalities and the patterns that can be discovered include description of the class or the concept (Characterization and Discrimination), Mining frequent patterns, associations, correlations, Classification and Regression for predictive analysis, Cluster analysis, Outlier analysis given by *Jiawei Han*.

Data mining is an integration of techniques from different domains like machine learning, databases and data warehousing systems, statistics, pattern recognition, and algorithms for visualization, information retrieval, high performance computing and many other domains.

Machine learning is a study of learning abilities of a computer and hence enhances the performance based on the data applied. It is a very fast growing discipline. There is high correlation between the problems in machine learning and data mining. Learning based on supervised, un-supervised and semi-supervised techniques are a few similarities between the machine learning problems and data mining problems.

Supervised learning is made possible with the help of labeled examples in the training dataset synonymous to classification problem. Unsupervised learning works with unlabeled examples in the training dataset synonymous to clustering technique. Semi-supervised learning makes use of both labeled and unlabeled examples when learning a model. Active learning enables users to play an active role in the learning process. Research in Machine learning puts its focus on the accuracy of the model. Prediction of the disease is one of the most interesting and challenging tasks where the development and application of data mining tools come into existence.

1.2 FEATURE SELECTION

Variable and feature selection play a major role in this research as this enables the user to have a clear analysis of the data. The three-fold objectives of variable selection includes the improvement of the predictive performance of the application, providing faster and cost-effective predictors and provide a better perspective of the underlying system under study. The third fold is the main crux of this research study. Unless one has a very clear understanding of the system being studied there would not be a possibility of in-depth study and analysis. Feature selection provides information which is in close proximity with the application being studied.

Feature selection plays a significant role in pattern recognition as it reduces the dimensionality by removing noise and irrelevant data as projected in [45]. This process enhances the performance of the mining task. The original representation of the variables is not altered through feature selection rather it selects a subset of the features which are more closely related to the problem domain. Given a set of X features the role of feature selection is to select a subset of these, x € X such that x is a subset of X.

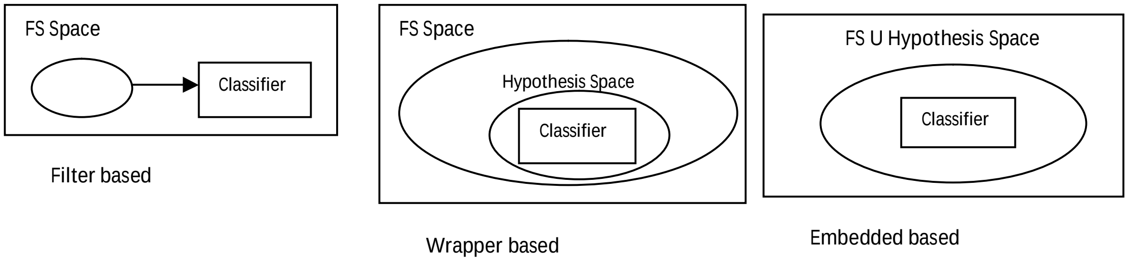

Figure 1.1: Feature Selection Techniques

Supervised and unsupervised learning use feature selection. Depending on how the search for feature selection is incorporated into the model being built, the techniques are classified into filter, wrapper and embedded methods.

Figure 1.1 gives an overview of the feature selection techniques. Filter based extracts the features and then trains the algorithm by inputting the subset into the model. Wrapper based builds the model along with feature selection. Embedded combines both filter and wrapper methods. Depending on the tenacity of the problem any one of the techniques can be applied for building the model. *Interestingness* is one measure used to rank and sort the attributes in the column. This interestingness measure gives information about the usefulness of an attribute.

Several algorithms have been applied in this research work to pick the efficient and suitable feature selection algorithm for the domain. Comparative analysis can be seen clearly in chapters 3 and 4. The purpose of applying feature selection techniques in the current research is to identify the most needed attributes thereby enhancing the performance of the classifier.

1.3 CLASSIFICATION TECHNIQUES

Classification algorithms model the categorical labels (discrete, unordered) while prediction algorithms model the continuous valued functions. Classification algorithms have been studied and applied in the current research like decision trees, support vector machine, neural networks. The main drive behind the use of the classification algorithms is the accuracy and efficiency of these algorithms. The categories of data are known and the algorithm should be in a position to predict the new incoming dataset.

The classifiers picked up in the current research have exhibited a very good performance on medical data, though they are termed as the traditional algorithms. Classification is an instance of supervised learning. The concrete implementation of an algorithm that implements classification is a classifier.

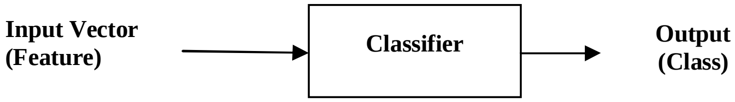

Figure 1.2: Working of a classifier

The term *classifier* indicates a mathematical function that maps input data onto a category (i.e., type of an output) as shown in figure 1.2 is represented mathematically as

$f = x \rightarrow y$

which can be written as

$y = f(x)$

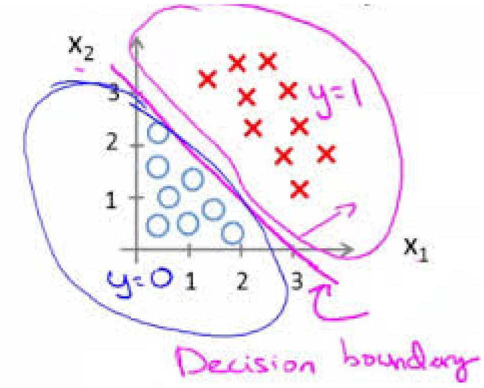

Figure 1.3: Decision regions formed by a classifier

A classifier classifies the input into decision regions as can be seen in figure 1.3. Decision boundaries or decision-region boundaries are formed between the decision regions. The input vectors are clouded with samples of different classes which the classifier is supposed to evaluate. An optimal classifier is one which produces least number of misclassifications.

The performance of a classifier depends vastly on the nature of the data being classified. No classifier works best on all given problems, performance is purely data-centric. *Precision* and *recall* are the popular metrics used for evaluating the performance of the classifier along with *receiver operating characteristic* (ROC) curves. Measuring the exactness or the quality of classifiers is precision whereas recall is a measure of the completeness or quantity of the classifier. High precision implies the classifier has picked more relevant than irrelevant results. High recall indicates the classifier predicted most of the relevant results.

Mathematical representations of the above metrics are as follows:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

where *tp* refers to *True Positives or true hits, fp* refers to *False Positives or false hits, fn* refers to *False Negatives or false errors*.

There are several types of classification algorithms categorized according to the purpose of use.

- Decision tree
- Rule approaches
- Probabilistic functions
- Networks

Each of these is elaborated in the coming sections.

1.3.1 Decision Tree Based Approaches

Decision trees are techniques used mostly because of their visualization and their understandability. In the medical domain these decision trees are used to identify the sequential nature of the attributes values and a decision that is based on these values. Decision trees are effective tools used extensively in medical domain. CART, ID3, C4.5 is a few popularly known and widely used decision tree algorithms. Splitting factor is different in each of these algorithms but among these, C4.5 has a better performance. These classifiers generate a decision tree from which the rules can be inferred. The rules in the form of *if...then* conditions improve human readability and hence have become widely used algorithms.

1.3.2 Rule based approaches

These rule based classifiers are highly expressive as the tree based algorithms are easy to interpret and also to generate. These help in classifying the new instances swiftly. The performances of these classifiers are comparable to that of the decision trees. The rules generated are mutually exclusive and exhaustive and contains as much information as the tree. These algorithms can easily handle missing data and numeric data. Learn One Rule (1R) extracts the best rule that covers the current set of training instances. Repeated

Incremental Pruning to Produce Error Reduction (RIPPER) and Prism algorithms are a few other algorithms. In the current research rule based approaches have not been studied.

1.3.3 Genetic Algorithms

Genetic Algorithms (GA's) have been designed based on the concepts of genetic mutation, modification, selection. This is based on the principles of Darwin's theory, evolution theory which is beyond the scope of this research and hence not applicable. They are applicable for optimization rather than for pattern recognition. They are used mostly in heuristic based search. They belong to the group of evolutionary algorithms used to solve optimization problems using techniques like inheritance, mutation, crossover etc. The possibility of the applicability of GA's in machine learning process is more if the number of rules to be applied is more. Research is being pursued in this direction as proposed in [28].

1.3.4 Support Vector Machine

It is a family of learning algorithms for classifying objects into their respective classes (two or more). Support Vector Machine (SVM) has been used in various domains specifically in medical diagnosis. The good performance of SVM in real-world applications is the means for its growing popularity. The algorithm is very robust towards high dimensional data. The sound theoretical base adds to these characteristics of SVM and thereby increasing the popularity of the algorithm in the field of machine learning.

Lot of research is still being undertaken to compare its efficiency with the traditional algorithms. It handles over-fitting of data by classifying the objects with large confidence. Hence SVM has been applied in the current research to compare its efficiency with the other algorithms. SVMs have been applied to different types of cancerous data and being used widely in the field of medical diagnosis.

1.3.5 Neural Network

Neural network (NN) or Artificial Neural Network is a "*computing system comprising of a number of simple and highly interconnected elements for processing, that* *can process the information by their dynamic state response to external inputs*" according to Dr Robert Hecht-Nielsen (Neural Network Primer: Part I). They have been widely recognized as powerful modeling tools and most vendors are embedding them into their data mining software. As universal approximators, NNs are most suitable for applications that need some regularities to be discovered or identifying associations.

NNs are expected to train the problem quite well compared to the conventional techniques, hence applicable to problems which are quite dynamic or non-linear in nature. They can extract meaning from complicated or imprecise data and thereby generate patterns for analysis and trends that the natural humans cannot perceive. Though the NN outperforms the other classification algorithms it has a disadvantage, it acts as a black box. The results produced from an NN cannot be interpreted by human experts.

Neural networks are useful for data mining and specifically for decision support applications. Human beings are experts in generalizing from experience whereas computers are good imitators of following instructions. Neural networks act as a bridge in a way by modeling the neural behavior of human brains on a computer. They are useful for pattern recognition or classification through a learning process.

Figure1.4 gives a view of the model of a neural network. As can be noted a neural network maps a set of input nodes onto a set of output nodes. The predictive accuracy of NN is high compared to the other methods or algorithms. Artificial Neural Network (ANNs) models are found to produce efficient computations and are considered as the universal approximators.

Neural networks process the information by simulating the human brain and hence enable high intelligence to the application; they learn by example. Layered feed forward networks were designed in 1950's. The limitation of a single layer network, they can solve only linearly separable problems, led to the development of Multi-Layer Perceptron (MLP). Varied types of NN models exist for predicting the accuracy, pattern classification.

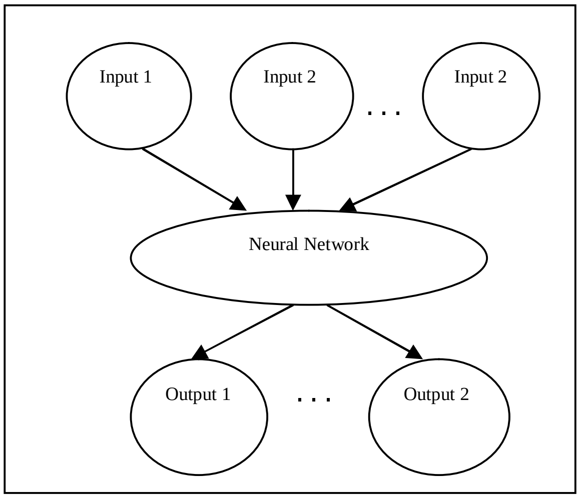

Figure 1.4: Model of a Neural Network

1.3.5.1 Feed forward Back propagation Neural Networks

Feed forward neural networks are one of the popular structures among neural networks. The main aim behind this network is the propagation of the inputs from the first layer i.e., the input layer to the immediate next layers till it reaches the output layer. The error is computed from the last layer i.e., the output layer and propagated back to the first layer in order to train the network. This process is repeated till the error is minimized.

A feed forward neural network is a collection of neurons that are inter-connected together in a network represented by a directed graph as shown in figure 1.5. The first layer of a feed forward network is the input layer and the last layer is the output layer with any number of hidden layers in between. The above figure shows a feed forward NN with 3 layers. Each neuron from a specific layer is connected to all the other neurons in the subsequent layer. The link between the **i**th and **j**th neurons is represented by the weight coefficient $w_{ij.}$ The degree of importance of the given link is based on the weight coefficient.

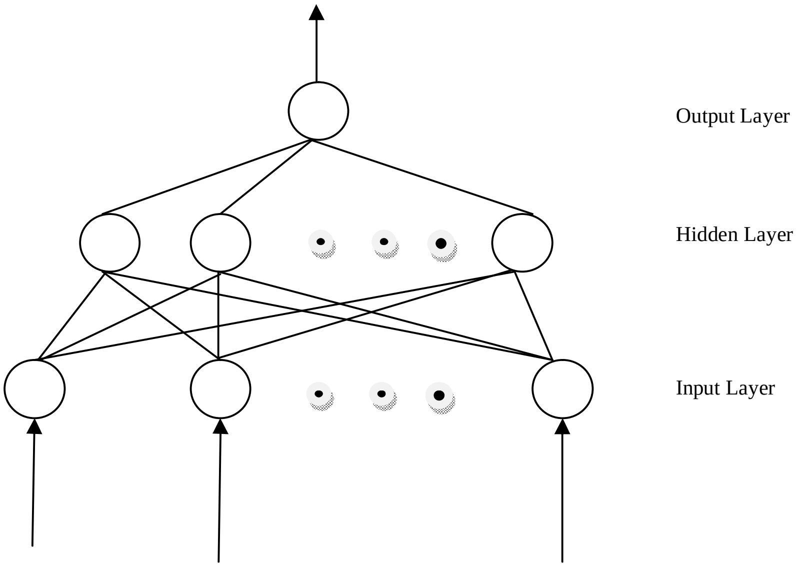

Figure 1.5: Feed forward neural network with three layers

Input layer takes in the input vector as represented in the dataset. The output of each input unit equals the corresponding component of the input vector. Each hidden unit computes the weighted sum of its inputs to form the scalar net activation. Net activation is the inner product of the inputs with the weights at the hidden unit, written as

$$net_j = \sum_{i=1}^{d} x_i \, w_{ji} + w_0$$

where the subscript i indexes the units in the hidden layer; w_{ji} denotes the input-to-hidden layer weights at the hidden unit j. The value of w_0 coming out is always 1. Each hidden unit gives an output which is a nonlinear function of its activation, f(net), that is

$y_j = f(net_j)$

The use of the back propagation algorithm is to train the neural networks used in conjunction with gradient descent. The gradient function is used to update the weights in order to reduce the loss of the function. This algorithm requires a desired output for the input value and hence is applicable to supervised learning method. Back propagation trains in two phases; propagation and weight update.

Back propagation networks are applicable to multi layer perceptrons and hence the multi layer network is supposed to have non linear activations for the multiple layers. MLP with feedforward back propagation has been applied in the current research as the performance of this algorithm is high compared to the other traditional algorithms.

1.3.5.2 Analysis of a Neural Network

The non linear modeling abilities and the capability of a neural network to learn, adapt and present human-like intelligence has enabled its recognition as a technique for machine learning which is powerful and more general. Neural networks are wonderful means for classifying and analyzing the data that belongs to any domain, specifically for medical diagnosis.

Inspite of all the power of an NN it still lacks the ability to provide explanation. There is no proper reasoning behind the learning system developed. As these systems are used for critical decision making process it would make sense if the decisions are understood and presented in human readable form. The presentation of these decisions in *if..then* form (human readable) would be a better way of projecting the output of the neural network model; since this enhances the modality of the decision.

The indecipherable nature of a neural network is mainly attributed to the knowledge of NN stored in the form of real valued parameters (weights and biases). Efforts are on by the research community in studying this network to generate human readable predicaments. Several algorithms have been designed which can be categorized as pedagogical, decompositional or eclectic approaches.

Decompositional approach develops local rules by disassembling the network architecture and then combines the local rules to generate the global rules for the complete architecture. Pedagogical approach extracts rules by analyzing the relationship between the inputs and outputs without disrupting the network architecture. Eclectic approach is a combination of the existing two approaches.

1.4 THESIS STATEMENT

The current research focuses on rule extraction mechanisms for feed forward back propagation neural network as this provides an enhanced solution to the black-box nature of a neural network or rather a hard to understand learning systems such as neural networks. The hypothesis proposed in this research is to build a rule extraction algorithm from neural networks which is simple, optimized and works even for small datasets accurately by using max-min approach.

1.5 THESIS OVERVIEW

The chapterization used in this research is as follows:

- Chapter 2 gives a summarized view of the various studies carried out in this area of research. Presents the literature study in a concise form.
- Chapter 3 presents the comparative study on different machine learning algorithms applied on the risk factors of cervical cancer. It concludes by picking up the most influential factors for the risk of cervical cancer.
- Chapter 4 presents a comprehensive study on the comparative analysis of machine learning algorithms applied on the stages of cervical cancer. The factors that lead to the different stages of cancer have been extracted and decision rules generated.
- Chapter 5 presents a novel algorithm for extracting rules from feed forward back propagation neural network algorithm. This is compared with the existing classification algorithms applied on the cervical cancer data. This algorithm is compared with the existing rule based approaches expounding the differences, similarities, limitations as such.
- The final chapter, chapter 6, gives an overview of the thesis, limitations, scope and further study in this work.