CHAPTER 2:

LITERATURE SURVEY

2.1. Machine learning

Machine learning is a branch of artificial intelligence that aims at solving real life engineering problems. It provides the opportunity to learn without being explicitly programmed and it is based on the concept of learning from data. It is so much ubiquitously used dozen a times a day that we may not even know it. The advantage of machine learning (ML) methods [1] is that it uses mathematical models, heuristic learning, knowledge acquisitions and decision trees for decision making. Thus, it provides controllability, observability and stability. It updates easily by adding a new patient’s record.

The application of machine learning models [2] on human disease diagnosis aids medical experts based on the symptoms at an early stage, even though some diseases exhibit similar symptoms. One of the important problems in multivariate techniques is to select relevant features from the available set of attributes [3]. The common feature selection techniques include wrapper subset evaluation, filtering and embedded models. Embedded models use classifiers to construct ensembles, the wrapper subset evaluation method provides ranks to features based on their importance and filter methods rank the features based on statistical measurements.
2.2. Computer aided diagnosis systems

A computer aided medical diagnosis system [4] generally consists of a knowledge base and a method for solving an intended problem. On the basis of the query posted to the system, it provides assistance to the physicians in diagnosing the patients accurately. The knowledge base of such medical systems relies on the inputs that spring up from the clinical experience of field experts. Knowledge acquisition is the process of transforming human expert knowledge and skills acquired through clinical practice to software. It is quite time consuming and labor intensive task. Common methods like Case Based Reasoning (CBR) solves the knowledge acquisition problem to some extent because the past records are maintained in a database, including possible remedies, past clinical decisions, preventive measures and expected diagnostic outcome measures. During patient diagnosis, the clinical database is matched for analogous past patient's record for taking suitable decisions.

Some of the major problems faced during the development of an expert diagnosis system are: medical experts are less interested to share their knowledge with others, experience knowledge (called common sense) is practically impossible to be separated and designing a unique expert system for diagnosing all diseases is difficult.

2.3. Software reliability

Software reliability [5] is defined as the probability that a system will not have a failure over a specified period of time under specific conditions. The knowledge of software reliability is very vital in critical systems because it indicates the design perfection [6]. In
this work, the primary aim is to enhance the software reliability of the computer aided diagnosis systems using machine learning algorithms.

To provide quality treatment and prevent misdiagnosis are the prime motivations for developing a medical diagnosis system. Diagnosing a disease of a patient accurately is a great challenge in medical field. A huge amount is spent on advanced primary health care devices based on software reliability research as they are considered as critical systems.

There are several software reliability models available in the literature; however, none of the models are perfect. An important research issue is choosing a suitable estimation model based on a specific application. One advantage of software reliability over hardware reliability is that a mechanical part surely undergoes ageing; suffer from wear and tear problem over time and usage; however software do not rust or wear out. Software reliability is a vital parameter for software quality, functionality and performance. Some common software reliability models are prediction and estimation models like bathtub curve, exponential, Putnam etc.

2.4. Supervised learning

Supervised learning is the most common form of machine learning scheme used in solving the engineering problems [7]. It can be thought as the most appropriate way of mapping a set of input variables with a set of output variables. The system learns to infer a function from a collection of labeled training data. The training dataset contains a set of
input features and several instance values for respective features. The predictive performance accuracy of a machine learning algorithm depends on the supervised learning scheme [8]. The aim of the inferred function may be to solve a regression or classification problem. There are several metrics used in the measurement of the learning task like accuracy, sensitivity, specificity, kappa value, area under the curve etc. In this work, the aim is to classify the patients as healthy or ill based on the past medical records. Before solving any engineering problem, it is vital that it is necessary to choose a suitable algorithm for the training purpose based on the type of the data. The selection of a method depends primarily on the type of the data as the field of machine learning is data driven. The next important aspect is the optimization of the chosen machine learning algorithms.

2.5. Classification task

Classification task [9] is a classical problem in the field of data mining which deals with assigning a pre-specified class to an unknown data. A learning model is built based on the relationship between the predictor attribute values and the value of the target [10]. The challenge is to correctly predict the class based on learning of past data. In machine learning, this kind of classification problems are referred to as supervised learning. Hence, we need to provide a data set containing instances with known classes and a test data set for which the class has to be determined. The success of the classification ability largely depends on the quality of data provided for learning and also the type of machine learning algorithm used [11]. For example, the classification techniques can be used to predict the fraud customers in a bank who apply for a loan or classify mangoes whether
they are good or bad and lots of other real time applications. The most common type of classification problem is binary classification, where the target has two possible values like good or bad, yes or no etc. There are several methods for measuring the classification performance like confusion matrix, lift curve, receiver operator characteristics etc.

2.6. Optimization

Every machine learning algorithm has a specific technique of learning and is based on the values of their parameters. When an algorithm is applied to solve a classification problem with a different set of parameters, the classification accuracy also differs abruptly in each case [12]. The challenge in machine learning to find the most suitable parameter values of the algorithms that solves an engineering problem to the best possible way in terms of performance metrics. Therefore, one has to fine tune the algorithm parameters that best suits the problem. There are several optimization techniques like genetic algorithm, particle swarm optimization [13], Tabu search methods etc. The focus of the study is to calibrate the algorithm parameters using design of experiment method.
References


