Relationship between machine learning and data mining

Education, Learning

Machine Learning has been used intensively and extensively by many organizations. It is becoming increasingly popular in healthcare as well. The large quantity of data collected and produced byhealthand human services are much complicated and copious to be refined and examined by conventional techniques.

Normally, machine learning utilizes data mining methods and some other learning algorithm to construct models of what is going on in the pool of data. It provides the methodology andtechnologyto transform these mounds of data into useful information for decision making or predicting future outcomes. This information is extracted through various data mining algorithms and techniques such as association, classification, clustering, and pattern recognition and is of great use to the medical experts.

DATA MINING

Customary physical data analysis is unproductive because the data has grown to Exabyte's. So there is a need of computer based study of this data. Though several approaches to computerized analysis of data are there, one of the approaches is data mining. It is a technique employed in extracting significant information from the available datasets. Data mining has become one of the most dominant areas of computerscienceand has its applications in almost every domain. Some of the application domains of data mining are:

- Disease Diagnosis
- Market Analysis
- Fraud Detection
- Customer Retention

- Science Exploration etc.

Data mining is an amalgamation of many technologies [2]. This is shown in fig. 1.

DATA MINING PROCESS

The process of data mining consists of various phases:

- Data collection: This is the first step of data mining process where the data is collected from various sources using specialized tools and techniques and is stored in a database for further processing.
- Feature extraction and data cleaning: This phase of data mining process involves two steps: Feature extraction and cleaning of data. The data collected is not always in a suitable form and must be transformed in a structure compatible with working algorithms. This is done using various feature extraction algorithms. The presence of noise, outliers, irrelevant data must be addressed i. e. the data must be cleaned before applied to any algorithm.
- Analytical processing and algorithms: This is the last phase of data mining process where analytical techniques are intended from the data processed.

APPLICATION OF DATA MINING IN HEALTH CARE SYSTEMS

Health care is diagnosing, preventing and providing better means for the treatment of various diseases. Currently bulk of data is generated in healthcare systems in the form of prescriptions, test reports, laboratory tests, billing reports [30]. This data is available in the form of datasets.

Dataset is an assembly of correlated data, distinct items of similar data which are arranged in the form of a database and can be accessed independently or in combination and is managed as a whole entity. Healthcare statistics systems tend to assemble this data in databases for research and analysis in order to assist in making medical diagnosis.

Many challenges are faced by our healthcare system and some of them are: heart diseases, Brest cancer, Brain tumor, DiabetesMellitus (DM), malarial infections etc. and all these must be diagnosed early with maximum accuracy.

Healthcare system deals with tremendous amount of data which is generated from various sources like hospitals, clinics, laboratories etc. while data mining can be implemented for retrieving the hidden patterns. It also assists to recognize the correlation between various patterns of medical data. Data Mining is one of the most significant practices in the area of healthcare for providing better diagnosis of any disease. Many tools, techniques and algorithms have been developed by researchers to facilitate the early diagnosis of a disease with maximum exactness.

MACHINE LEARNING

The ability of computers to study from experiences without being explicitly programmed is machine learning. The traditional form of programming involves coding all the rules and machine will generate output based on the logical statement i. e. the rules. However when the system becomes complex, more rules are needed to be written and hence becomes unsuitable to retain.

However machine learning is thought to overcome this issue. Here the machines learn how the input and output data are correlated and then accordingly writes a rule. The programmer doesn't need to write new rules every time a new data or situation arises. Learning and inference are the principal objective of machine learning.

HOW TO IMPLEMENT MACHINE LEARNING

The following steps define the implementation of machine learning.

1. Define a question.
2. Collect the data relevant to that question.
3. Visualize the data.
4. Train the algorithm.
5. Test the algorithm.
6. Collect the feedback.
7. Refine the algorithm.
8. Loop 4-7 until results are satisfying

The diagnosis is again confirmed by assessing the thyroid hormone levels in blood which get increased (T3 and T4) and the TSH levels get low. A thyroid scan is also often done and corroborates the diagnosis and can establish inflammation as a cause.

LITERATURE REVIEW

In order to address the issue of diabetes and its early diagnosis, it becomes essential that previous work in the said pursuit be recorded to ease further research/progress. So in this section we review the related work regarding thyroid disorder prediction through various classification techniques.

The diagnosis of any human abnormality is mainly driven by a Computer Aided Diagnosis (CADx) framework. This framework mainly encapsulates 3 modules: Input data, Extraction of features/Selection of best features and Classification. Based on the conventional CADx framework, we have made an attempt to review previous work based on the following parameters viz: Extraction technique used, Feature selection procedures used (in some studies), Classification algorithms used and finally the accuracy of each proposed work.

M. R. Nazari Kousarrizi et al [31] stated that Hyperthyroidism (an overactive thyroid) may also get induced by inflammation of thyroid gland, several types of medical drugs, and lack of control of yield of hormone of thyroid. The onset of disorder of thyroid gland should not be ignored as or underestimated as severe hyperthyroidism, also known as thyroid storm, and last stage of hyperthyroidism, known as myxedema coma, may lead to death of a person [32, 34].

They also generalized that Diagnosis of Thyroid disorder is an essential classification problem. Sound rendering of the thyroid data is significant probe in the thyroid disease diagnosis [32, 33, 35] Several new techniques, like genetic algorithm, SVM, ANN, decision trees etc., have been applied to put patients into properly defined status, position or condition [32, 33, 35, 36]. The researchers also proposed method that has two stages. In the primary stage, feature selection has been used as a preliminary processing step.

The principle aim of feature selection is to cut the count of features used in classifying the subjects while maintaining satisfactory classification accuracy [37]. In this research sequential backward selection, Genetic Algorithm and sequential forward selection have been used as feature selection techniques. In the following stage, Support Vector Machine is utilized in classification of objects. The researchers debate that the chosen features received from the proposed technique are same as clinical experimentations utilized by medical specialists to diagnose the disease of thyroid.

Genetic algorithm (GA) is a class of optimization methods enlivened by the biological process of replication. It has been utilized to resolve several jobs including object recognition [42], target identification [43, 44], facial recognition [45, 46]. Genetic Algorithm uses series of iterations on a universe of structures, each one of which constitutes a candidate solution to the problem at present, specifically converted (encoded) in form of string of typifies. An arbitrarily produced set of such chains of typifies or symbols form the initial pool of solutions from which the Genetic algorithm starts its searching job.

There are 3 primary genetic operators which direct this searching task namely selection, crossover & mutation. This searching process is iterative in which every string is selected, evaluated and re-combed until some terminal state is reached. Furthermore, evaluation of solution string of symbols is dependent on a fitness function which is in-turn is problem-dependent.

Support vector machines: In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyses data and recognize patterns, used for classification and regression analysis [40]. The SVM is based on statistical learning theory. The SVM solves the problem of interest indirectly, without solving the difficult problems. The support vector machine presents a partial solution to the bias variance trade-off dilemma.

Decision tree: A decision tree is a predictive modelling technique used in classification and prediction tasks [41]. Decision tree uses divide and conquer technique to split the problem search space into subsets. A decision tree is a tree where the root and each internal node is labelled with a question. The arcs emanating from each node represents each possible answer associated with question. Each leaf node represents a prediction of a solution to the problem under consideration.

Astha Rastogi , Monika Bhalla [38] stated An Artificial Neural Network (ANN) is a system of interlinked computing constituents namely nodes. Data is processed by the fundamental interaction amongst nodes. Knowledge is not present in these processing nodes; rather it is represented by the weights of the associations or connections between these processing nodes.

In simplest form of neural network, known as feed forward network, information or entropy moves only in one direction i. e. in forward direction, from the input processing nodes to the output processing nodes through the hidden layer or layers of nodes. Since data move in one direction, no cycles or loops in the network can exist. External or user information gets in the network via the input layer of processing nodes while the output layer of nodes produces the model situation or result. The hidden layer of nodes furnishes connections or associations necessary to discover complex patterns hidden in the data. [39]

Ozyilmaz et al. [14] in 2002 used various neural network methods including Multi-Layer Perception with Back-Propagation method (MLP), Radial Basis Function (RBF) and adaptive Conic Section Function Neural Network (CSFNN) to help diagnosis of disease of thyroid, their classification and categorization precisions are respectively 88%, 81% and 85%. In 1997, Probabilistic Potential Function Neural Network (PPFNN) claxssifier [15] was employed and the accuracy of 78. 14% was obtained.

Pasiet al. [16], In 2004, applied five different methods including Linear Discriminant Analysis (LDA), C4. 5 with default learning parameters (C4. 5-1), C4. 5 with parameter c equal to 5 (C4. 5-2), C4. 5 with parameter c equal to 95 (C4. 5-3) and DIMLP with tw M. R. Nazaro hidden layers and default learning parameters (DIMLP) to perform classification, and the accuracies reached 81. 34%, 93. 26%, 92. 81%, 92. 94% and 94. 86% respectively.,

Polat et al. [17], In 2006, proposed application of artificial immune recognition system (AIRS) with an accuracy of 81%. Furthermore, the author studied a hybrid method that combines AIRS with a developed Fuzzy weighted pre-processing, and obtained a classification accuracy of 85%. In 2008, Keles et al. [18] diagnosed thyroid diseases with an expert system that called ESTDD (expert system for thyroid dis- ease diagnosis), whose accuracy was 95. 33%. In 2009, Temurtas [19] realized the diagnosis by Multi-Layer Perception with Levenberg-Marquardt (LM) algorithm (MLP with LM), and the corresponding accuracy was 93. 19%.

In 2011, a Generalized Discriminant Analysis (GDA) and Wavelet Support Vector Machine (WSVM) (GDA-WSVM) [20] method for diagnosis of thyroid diseases was presented, and obtained 91. 86% classification accuracy.
In 2011, Chen [21] proposed a particle swarm optimization optimized support vector machines with fisher score (FS-PSO-SVM) CAD system for thyroid disease, and the average accuracy of 97. 49% was achieved.

Arpneek Kaur et al. have used Multilayer back propagation network and self-organized map for diagnosis of thyroid disease. Weka tool has been used for the purpose. Thyroid data set is used for predicting the disease. The performance of BPN and SOM networks were found by varying the number of neurons present in the hidden layer of the network and also by the percentage of training data [22]. Yilmaz Kaya et. al. Have developed an Extreme Learning Model which is a single hidden layer feed forward neural network and is trained with gradient based learning algorithm.

This extreme learning model was used in the diagnosis of thyroid disorders by performing classification and the accuracy of the classification was found to be 96. 75%. 70% of the samples were used for training and 30% samples were used for testing purpose [23].

Rajkumar Nallamuth et al. have formulated several classification frameworks for diagnosis of the disease of thyroid. The classification models include C4. 5, Multilayer Perceptron, and radial basis function networks. It is observed that MLP has performed well compared to other classifiers [24]. Md. Dendi

Maysanjaya et al. have used MLP model using back propagation algorithm and the results of which have been compared with WEKA tool and RBF Network. It is found that MLP and RBF have given much accuracy in predicting the thyroid disease [25].

The authors Kenji Hoshi and Junko Kawakami have developed a Bayesian regularized neural network and a self-organized map to predict hyperthyroid and hypothyroid using linear discriminate analysis. For this, they have used hormones related thyroid predicting the disease [26].

Similarly Jasdeep singh bhalla and Anmol agarwal have developed hybrid neural networks for medical diagnosis using scaled conjugate gradient back-propagation and Marquardt back propagation algorithm. The input to the model is a thyroid dataset and basing on this dataset they have compared the performance of the models and found the efficiency of the prediction of ANN’s in medical diagnosis [27].

Hasan makas et al. developed seven distinct sorts of Neural Networks keeping in mind the end goal to recognize more strong and dependable systems for diagnosing the thyroid illness. They have utilized swarm optimization and ant bee colony algorithm for training and testing the networks for diagnosis of thyroid disease [28].

Chang et al adopted seven feature extraction technique viz., co-occurrence matrix, grey level run-length matrix, lawstextures energy measures, wavelet and Fourier features based on local Fourier coefficients[47].

Senol et al proposed ahybrid structure in which Neural Networks and Fuzzy logic are combined to diagnose the thyroid disease [48].

Rouhani et al compared several ANN models for diagnosing the thyroid disease [49]. Isa et al has experimented for several activation functions such as sigmoid, Hyperbolic tangent, Neuronal, Logarithmic and Sine activation function for the MLP Neural Network and determined the most suitable function to classify the thyroid disease as Hypothyroid and Hyperthyroid[50].

Prerana, Parveen Sehgal et al [51], proposed a precise technique for detecting the thyroid by utilizing the back propagation algorithm. Artificial Neural Network is developed using the back propagation of error to identify the preliminary thyroid prediction. ANN is trained subsequently for testing the experimentally, but not the same training sets.

The training can be done in two ways as supervised learning and unsupervised learning. The experimental result is carried out in MATLAB Neural Network Toolbox Software. This provides better performance than the simple gradient descent algorithm.

3. Aims and objectives:

There is scarcity of medical resources especially in developing countries. Knowledge discovery from thyroid disorder data sources can become one of the main sources for thyroid disorder detection. This research aims to diagnose thyroid disorder using data related to thyroid disorders and deep learning algorithms. Therefore the variousgoalsof this research are:

- To study and analyse the existing classifiers and feature selection techniques for predicting classification accuracy of Thyroid disorders.
- Evaluation: various data mining techniques on data related to thyroid disorders for comparison purposes.
- To implement feature selection technique on the proposed ensemble model for improving performance in classification of Thyroid disorders.
- Propose a predictive model based on knowledge discovery mechanism using data related to thyroid disorders and deep learning algorithms, which will be used in thyroid disorder evaluation.
- To compare results of implemented work with existing research for better performance in terms of standard metrics.

METHODOLOGY

Methodology refers to the steps involved in carrying out the research work. The aim of this research is to enhance the classification accuracy of thyroid disease prediction. This research is carried out in the following phases:

Phase 1: In the first phase of our research following work is carried out:

- Dataset is chosen: Different healthcare datasets are available and we have chosen a Thyroid dataset.

- Data is pre-processed: Pre-processing refers to a data mining method used to perform necessary processing of data. It may include missing data, noisy data and inconsistent data
- Model evaluation: Existing Machine learning models with feature selection techniques will be tested. After this an ensemble model will be prepared using these machine learning algorithms, and its performance will be evaluated.

Phase 2: In this step a feature selection technique will be developed using various feature selection algorithms.

- Proposed technique will be implemented for the ensemble model to optimize classification accuracy.
- The results of implemented work will be compared with existing ones for better performance.

Phase 3: This is the last phase of our research and the following work is carried out:

- Resampling the Dataset: In this step a resampling technique will be implemented on our dataset to avoid the problem of class imbalance.
- Implementing the feature selection: The proposed feature selection technique will be implemented on the balanced dataset (resampled dataset).
- Training the model: The model is now trained using this dataset.
- Result evaluation: The results of implemented work will be compared with existing ones for better performance. The performance is determined by various metrics.