Machine Learning Algorithms for Disease Diagnosis Literature Review

K.Vidhya, M.Vanathi, P.Vinukirthi, P.J.Swetha

KPR Institute of Engineering and Technology, Coimbatore ,Tamilnadu , India.

Abstract- **In medical field, computerized disease diagnosis is a rapidly growing dynamic area of research. In recent years, important attempts are made for the improvement of computerized diagnosis of disease because errors in medical diagnostic systems can result in seriously misleading medical treatments. Machine learning is important in data analysis and prediction .An effective prediction from already available data requires proper support of machine learning techniques. Machine learning techniques basically supportable classification of data and for predicting result. In the field of bio-medical, pattern recognition and machine learning promise the improved accuracy of perception and diagnosis of disease. They also promote the objectivity of decision-making process. For the analysis of high-dimensional and multimodal bio-medical data, machine learning offers a worthy approach for making classy and automatic algorithms. This survey paper provides the details of different machine learning algorithms.**

I. INTRODUCTION

In today's society there are large amount of data. For every transaction there is an accomplishment stored somewhere. We can get more amount of data than we hope to analyze. The natural hope is the analysis of data can be done by computer.Data mining refers to the varied tasks of analyzing keep information for patterns seeking clusters, trends, predictors, and patterns in an exceedingly mass of stored information. For an example, several grocery stores currently have client cards, rewarding frequent users of the grocery with discounts on explicit things. The stores provides these cards to encourage client loyalty and to gather information for, with this cards, they'll track a client between visits to the shop and doubtless mine the collected information to work out patterns of client behaviour. This is often helpful for deciding what to market through promotion, shelf placement, or discounts. An outsized banking corporation makes many choices regarding whether or not to just accept a application or not. It may have many kind of data regarding the individual age, employment history, salary, credit history and regarding the application quantity, purpose, rate. In addition, the bank has the same information about thousands of past loans, plus whether the loan proved to be a good investment or not.From these conditions, the bank wants to know whether it should make the loan. Data mining can potentially improve the loan acceptance rate without sacrificing on the default rate, profiting both the bank and its customers. Artificial Intelligence can enable the computer to think. Computer is made much more intelligent by AI. Machine learning is the subfield of AI study. Many researchers think that without learning, intelligence cannot be developed. There are numerous types of Machine Learning Techniques that are shown in Fig.1 Supervised, Unsupervised, Semi Supervised, Reinforcement and Evolutionary Learning.

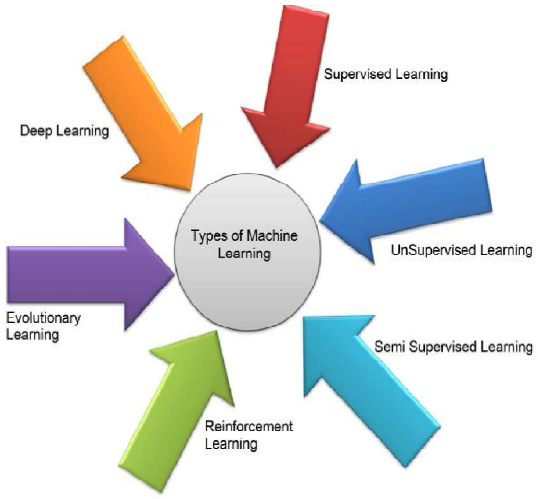

Fig.1 Types of machine learning techniques

Deep Learning is one of the types of machine learning techniques. These approaches are used to classify the data set.

A. Supervised Learning

Offered a set of examples with suitable targets for training and on the basis of this training set, algorithms respond properly to any or all possible inputs. Learning from exemplars is another name of Supervised Learning. Classification and regression are the kinds of Supervised Learning. Classification: It provides the prediction of Yes or No, for example, “Is this tumor cancerous?”, “Does this cookie meet our quality standards?” Regression: It provides the solution of “How much” and “How many”.

B. Unsupervised Learning

Correct responses or targets don't seem to be provided. Unsupervised learning technique tries to seek out out the similarities between the input data and supported these similarities, un-supervised learning technique also classifies

the data. This is also called as density estimation. Unsupervised learning contains clustering [1] which makes clusters on the premise of similarity.

C. Semisupervised Learning

Semi supervised learning technique may be a category of supervised learning techniques. This learning additionally used to unable the data for training purpose (generally a minimum quantity of labeled-data with a large quantity of unlabeled-data).Semi-supervised learning also lies between unsupervised-learning (unlabeled-data) and supervised learning (labeled-data).

D. Reinforcement Learning

This learning is encouraged by behaviorist psychology. Algorithm only informs about our answer whether it is right or wrong but it does not mention the ways of correcting it. It is like learning with the critic as the algorithm explores and tests various possibilities until it finds the right one. It does not recommend improvements. Reinforcement learning is different from supervised learn- M. Fatima, M. Pasha 3 ing in the sense that accurate input and output sets are not offered, nor suboptimal actions clearly précised. Moreover, it focuses on on-line performance.

E. Evolutionary Learning

As we consider our learning process to be biologically evolutional, progress are made about their survival rates and chances of having off springs by adapting the biological organisms. This model can be used in a computer to measure the accuracy of the solution by incorporating the idea of fitness.

F. Deep Learning

This learning uses a set of algorithms. These algorithms have high level abstraction on the data. Various linear and nonlinear transformation processing layers are in deep graph which is utilized by this learning algorithm.

II. MACHINE LEARNING ALGORITHMS

One of the main-stays of information technology is Machine learning. Increasing amount of data in every domain becoming available worldwide made a good reason to believe that a necessary ingredient in technological areas is to make smart data analysis .Machine learning solves as many problems by also providing the guarantee for the solution [2].Fig.2 shows the process of machine learning .

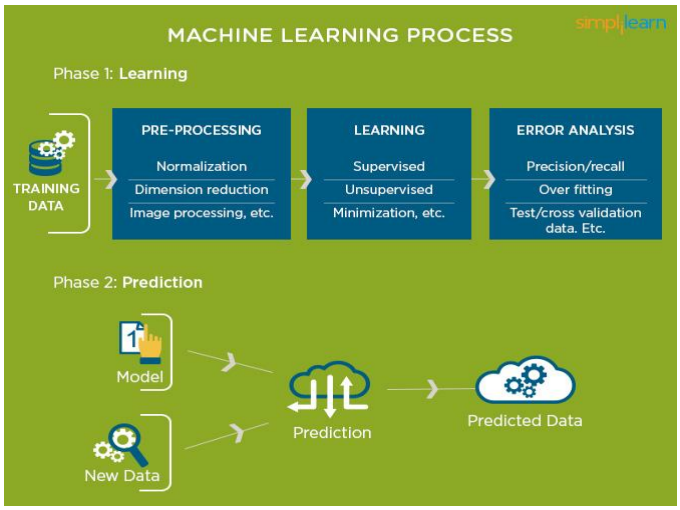

Fig.2 Process of Machine Learning

A. SVM

One such supervised learning models that analyses data after which they uses for classification purpose is Support Vector Machine. Classification here refers to relating the images to a particular class or to a dataset or for a set of categories . It is the standard set of supervised learning model with the aim to find the best highest-margin separating hyperplane given a two class training sample and those hyperplane should not lie closer to other class for better generalization[3] . The one that is far from data points from each category should be selected as a Hyperplane and the points lying near to the margin of the classifier are the support vectors[4] .

Classification an instance of supervised learning is considered as a task of inferring a function from labelled training data in machine learning. The correctly identified images are considered as training data in image retrieval process as that are placed in an particular class and each class belong to different category of images. The new examples in this algorithmic model are assigned to one category class or other. This representation of examples in categories are with clear gaps that are as vast as possible[5].

B. Naive Bayes Classifier

A classification technique, Naive Bayes, with a notion defines all features are independent and unrelated to each other. It also defines that,a status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is a powerful algorithm employed for classification purpose. It works well with imbalancing problems and missing values. Naive Bayes employs the Bayes Theorem. Using Bayes theorem posterior probability P(C|X) can be calculated from P(C),P(X) and P(X|C) [R6]. Therefore, P(C|X)=(P(X|C) P(C))/P(X)

Where,

P(C|X) represents target class's posterior probability .

P(X|C) represents predictor class's probability.

P(C) represents class C's probability being true.

P(X) represents predictor's prior probability.

C. Decision Tree Classifier

Decision tree is a category of supervised machine learning algorithm used to solve classification problems. It's main objective is to predict the target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features and it can have two or more branches while the leaf nodes represent classification. Decision tree evaluates the highest information gain among all the attributes for choosing the nodes in every stage[7].

D. Artificial Neural Network

Artificial Neural Network, an interconnected group of artificial neurons, uses a mathematical model based on the connectionist approach for computation. The connection site between neurons is Synapse. In Artificial Neural Network (ANN), a continuous variable xj replaces the original spikes(short pulses of electricity).The rates xj of all neurons which send signals to neurons i are weighted with parameters wij .Here these weights are describing the efficacy of the connection from j to i.e. weight is therefore called as 'synaptic efficacy'. The output xi of neuron i will be a nonlinear transform of the summed input, where n represents the number of input lines converging onto neuron i and # represents a formal threshold parameter. Neuron has three basic parameters -weights, threshold and a single activation function [8].

E.K-Means Algorithm

K-mean is a clustering algorithm and it partitions the given n observations into K clusters. The mean of each cluster is found and the image is placed in an cluster, having a mean with the least Euclidean distance .The k-mean clustering cannot separate images due to the complex distribution with different concepts. Clustering like regression describes the class of problem and methods. Clustering methods has two modelling approaches like Centroid-based and Hierarchical. The most popular among all unsupervised learning is K-mean which solves the clustering problem. Its procedure follows a simple way to classify a given data set through a certain number of clusters (take as K clusters).Data points are homogeneous inside a cluster ,and ,to peer groups they are heterogeneous.

Clusters in k-means have its own centroid. Sum of square value for a cluster is the sum of square of difference between centroid and data points within a cluster. The sum of squares values for all the clusters is the total within the sum of square values for the cluster solution. As the cluster number increases ,this value keeps on decreasing but if we plot the result we can see the sum of squared distance decreases to the same value of K , after that its slows down we can find the optimum number of clusters [R9].

F. Multiple Linear Regressions

A statistical model that describes data for explaining the relationship between one dependent variable and two or more independent variables can be done by Multiple Linear Regression. Analyzing the correlation and fitting the line, evaluating the validity, directionality of the data, and usefulness of the model are the different stages of multiple linear regression model. For a given combination of the input factors the regression line represents the estimated disease chance. Scatter plot is defined by a linear equation of $y=\beta 0 + \beta 1\ x1+ \beta 2x2+.....+ \beta ixn$ for i = 1…n. The deviation occurring between regression line and the single data point is the variation that our model cannot explain is known as residual[10].

G. Principle Component Analysis

Many applications like face recognition, pattern recognition, image compression and data mining uses a statistical technique known as Principle component analysis (PCA).It acts as multiple factor analysis and as correspondence analysis for handling heterogeneous sets of variables and quantitative variables respectively. Mathematically PCA depends on Singular Value Decomposition (SVD) of rectangular matrices and Eigen decomposition of positive semi definite matrices. The size of a dataset is reduced by retaining maximum of information about original dataset. A mathematical procedure is included to convert a large set of correlated variables into a smaller set of uncorrelated variables called principal components (PCs) . The problem can also be stated as follows: Find a lower dimensional representation of it , y = (y1, ..., yD) T with D, Given the d-dimensional random variable x = (x1, ...,xd) T.Thus, the input to HMV ensemble model is a dimensionally reduced dataset that classifies and predicts diseases providing better results than the traditional approach which improves the accuracy, reduces the noise and irrelevant data[11].

H. Ensemble Methods

Ensemble methods are one such meta-algorithms that combine several machine learning algorithms and techniques into one predictive model to decrease the variance, bias or improve the predictions. The sequential ensemble methods are derived totally from the base learners and primary motivation is mainly to exploit the dependence that falls in between the base learners. The overall performance is increased and boosted by weighing all the previously mislabeled examples with higher weight. There is also parallel ensemble methods where the base learners are generated in parallel (e.g. Random Forest) and it exploits independence that falls in between the base learners since the error here can be reduced dramatically by averaging. Most of the ensemble methods make use of a

single base learning algorithm to produce learners who fall in the same type, leading to homogeneous ensembles. There are also some methods that are continuously using learners that are of different types, this leads to heterogeneous ensembles. In order to be more accurate than any of its individual members, the base learners should have to be as accurate and diverse as possible.

III. CONCLUSION

The machine learning techniques are being widely used to solve real world problems in these days by storing, manipulating, extracting and retrieving data from large sources. Supervised machine learning approaches are widely adopted and these techniques are proved to be very expensive when the systems are implemented over wide range of data. This is due to the fact that significant amount of effort and cost is involved because of obtaining large labelled datasets. Thus active learning provides away to reduce the labelling costs by labelling only the most useful instances for learning. An unsupervised learning is also used in summarization and association analysis. The machine learning types and the algorithms are explained in this survey. Furthermore, there are more and more techniques that apply machine learning as a solution .

REFERENCES

[1]. Meherwar Fatima1, Maruf Pasha, Survey of machine learning algorithm for disease diagnostic Journal of Intelligent Learning Systems and Applications, 2017, 9, 1-16

[2]. A. Smola and S. Vishwanathan, INTRODUCTION TO MACHINE LEARNING. United Kingdom at the University Press, Cambridge, October 1, 2010

[3]. Isodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010 , 554–559doi:10.1109/CICN.2010.109.

[4]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038

[5]. Sunpreet Kaur, Sonika Jindal, A Survey on Machine Learning Algorithms, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2763 Issue 11, Volume 3 (November 2016)

[6]. Rish, I., 2001. An empirical study of the naive Bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM.pp. 41–46.

[7]. Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining &Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101,arXiv:1502.03774

[8]. W. Gerstner, Supervised learning for neural networks: a tutorial with exercises

[9]. K. M. M. Y. Dietterich T.G., “Applying the weak learning framework to understand and improve c4.5,” no. pp 96-104, san francisco:morgan

[10]. Niti Guru, Anil Dahiya, Navin Rajpal, “Decision Support System for Heart Disease Diagnosis Using Neural Network”, Delhi Business Review, Vol.8, No.1, 2007

[11]. Ian T. Jolliffe, Jorge Cadima: Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society, A Mathematical, Physical and Engineering Sciences, 13 April 2016 Volume 374, issue 2065.