International Journal of Pure and Applied Mathematics
Volume 117 No. 16 2017, 79-85
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

A REVIEW ON DISEASE DIAGNOSIS USING MACHINE LEARNING TECHNIQUES

[1]Shaik Razia, [2]P.Swathi Prathyusha,[3]N.Vamsi Krishna,[4]N.Sathya Sumana
[1]Assistant Professor, CSE Department, K L University, Vaddeswaram, Guntur, Andhra Pradesh, India.
[2, 3, 4] Student, CSE Department, K L University, Vaddeswaram, Guntur, Andhra Pradesh, India.
[1] razia28sk@gmail.com,[2] swathiprathyusha444@gmail.com, [3]vamsikrishnanori97@gmail.com

Abstract: In Disease Diagnosis recognition of patterns is so important for identifying the disease accurately. Machine learning is the field which is used for building the models that can predict the output based upon the inputs which are correlated based upon the previous data. Disease identification is the most crucial task for treating any disease. Classification algorithms are used for classifying the disease. There are several classification algorithms and dimensionality reduction algorithms used. Machine Learning gives the PCs the capacity to learn without being modified externally. By using the Classification Algorithm a hypothesis can be selected from the set of alternatives the best fits a set of observations. Machine Learning is used for the high-dimensional and the multi-dimensional data. Classy and automatic algorithms can be developed using Machine Learning.

Keywords: Machine learning, classification algorithms, Decision trees, KNN, K-means, ANN

1. Introduction

Disease diagnosis is abbreviated as Dx or Ds. This is the process of determining which disease explains a person's symptoms. Many signs and symptoms are non-specific and hence the diagnosis is the most challenging job. We can do the disease diagnosis using Machine Learning techniques.
We can develop a model in which the user can enter his symptoms and the model gives a particular disease. Machine Learning gives the PCs the capacity to learn without being modified externally.. There are many types of Machine Learning:

Supervised:

It can be seen as a Machine Learning job of concluding a function from named training information. The training information will have an arrangement of preparing cases in which every instance is a combination of input object(typically a vector) and a required yield value(also called as supervisory flag).

Unsupervised:

It can be seen as a Machine Learning job used to draw derivations from datasets which contains input information without named responses. Cluster analysis is the most widely recognized unsupervised Learning technique.

This technique is utilized for data examination to discover designs which are unseen.

Deep Learning:

This is likewise called as deep structured Learning or various leveled Learning. It is a piece of more extensive group of Machine Learning strategies which depend on learning information representations, rather than specific algorithms.

Semi Supervised:

This learning strategy is the class of supervised learning procedures. This learning method utilizes unlabeled information for training reason. Semi supervised learning procedure lies in the middle of the supervised learning which utilizes the named information and the unsupervised learning which utilizes the unnamed information since it for the most part utilizes the base measure of labeled information with a tremendous measure of unlabeled information.

Reinforcement:

This learning advises the algorithm when the appropriate response isn't right yet doesnot give a procedure in which it can be revised. It needs to test different potential outcomes until the point when it finds the correct one.

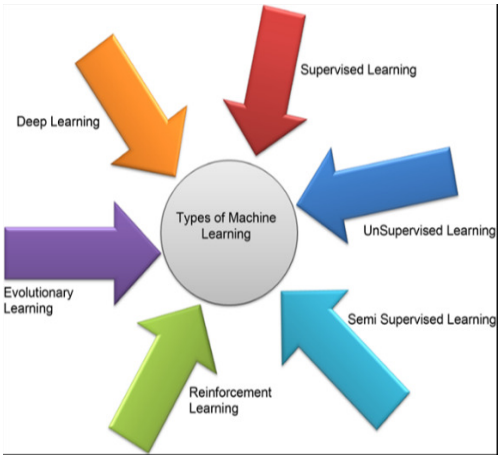

Figure 1. Types of Machine Learning

2. Literature Survey

In 2013 the researchers Vikas Tiwari and T.D.Diwan[1] (Vikas Tiwari, 2013) has given a paper that gives an automatic and hidden approach to identify, patterns that are hidden, of cancer disease. The given system use data mining techniques such as association rules and clustering. The methods involved in the data mining techniques are data collection, data processing, categorization of data set and rule mining. Attribute based clustering for feature selection is an important task of this paper. In this method we use vertical fragmentation in the data set. Here the data set is divided into two clusters, one cluster has all the relevant attributes and the other cluster has all the irrelevant attributes.

In 2006 a researcher M.peleg,S.tu[2] (M.peleg, 2006) has given a paper named Decision Support ,Knowledge Representation and Management. The clinical decision support is complete program designed to help the health professionals in making clinical decisions. The system has been considered as an active knowledge system. The main objective of the modern clinical system is to assist clinicians at the point of care. The objective of the system is to give the needed information with the health care's organizational dynamics. Decision support systems are implemented by standardization in information system infrastructure. The system give sits support in the complex tasks of differential diagnosis and the therapy planning. The system has to work on the knowledge modeling task in which modelers gives the medical knowledge that enables the system to deliver appropriate decision support system. The developers of the above system have two knowledge management tasks, one is the project oriented tasks that elucidate the organizational goals, responsibilities and the other is the communication and the co- ordination patterns of the care process in which the system has to operate.

In August 2013 a researcher Mohammed Abdul Khaleel has given a paper which is a Survey of on Medical Data for Finding Frequent Diseases using data mining techniques[3] (Khaleel, 2013) . This paper concentrate on examining data mining strategies which are required for medical information mining particularly to find frequent infections, for example, heart sicknesses, lung malignancy, breast cancer etc. Data mining is the process of extracting data for discovering latent patterns which can be translated into valuable information. The data mining techniques have being applied to medical data include Apriori and FPGrowth and unsupervised neural networks, linear programming, Association rule mining. The association rule mining discovers frequently occurring items in the give dataset. The medical mining yields required business intelligence to support well informed diagnosis and decisions.

A researcher Vembandasam yet al. [4] (Vembandasamy.K, 2015) played out a work, to analyze coronary illness. In this the algorithm utilized was Naive Bayes algorithm. In Naïve Bayes algorithm Bayes hypothesis is utilized. Henceforth, Naive Bayes has an effective freedom presumption. The utilized data collection is gotten from a standout amongst the most driving diabetic research organizations in Chennai, Tamilnadu. There are more than 500 patients in the dataset. The device utilized is Weka and classification is executed by utilizing 70% of Percentage Split. The exactness offered by Naive Bayes is 86.419%.

The data mining approaches are suggested to be applied by the researchers Chaurasia and Pal [5] (Chaurasia.V, 2013) to detect heart disease. An information mining device WEKA is utilized which contains an arrangement of machine learning algorithms with the end goal of mining. For this viewpoint Naive Bayes, J48 and bagging are utilized. Coronary illness informational set is given by UCI machine learning lab that comprises of 76 traits. For expectation just 11 attributes are used.82.31% precision is given by Naïve Bayes. J48 gives 84.35% of rightness. 85.03% of exactness is accomplished by Bagging. Bagging gives the better classification factor on this informational set.

The researchers Parthiban and Srivatsa [6] (Parthiban.G, 2012) have done a work on finding of coronary illness in diabetic patients. For this machine learning techniques are utilized. Naive Bayes and SVM methods are connected by utilizing WEKA. There are 500 patients in the informational set which is gathered from Research Institute of Chennai, Tamilnadu. There are 142 patients who have infection and there are 358

patients whose disease is missing.74% of exactness is accomplished by utilizing Naïve Bayes Algorithm. The most elevated exactness of 94.60 is achived by utilizing SVM.

A researcher Tan et al. [7] (Tan.K.C, 2009) has given a crossover strategy. In this two machine-learning methods like Genetic Algorithm (G.A) and Support Vector Machine (SVM) are consolidated viably. This is finished by wrapper approach. The tools utilized as a part of this investigation are LIBSVM and WEKA. For this investigation five informational sets like Iris, Diabetes infection, breast Cancer , Heart and Hepatitis ailment are taken from UC Irvine machine learning storehouse. After applying GA and SVM cross breed approach, exactness of 84.07% is accomplished for heart disease.78.26% precision is acquired for diabetes information set.76.20% exactness is accomplished for Breast growth. Precision of 86.12% is accomplished for sickness of hepatitis.

A researcher Iyer et al. [8] (Iyer.A, 2015) has done a work to foresee diabetes illness by utilizing two methods decision tree and Naive Bayes. Ailments can happen when insulin production is inadequate or the utilization of insulin is inappropriate. Pima Indian diabetes informational set is utilized as the informational set in this work. Utilizing WEKA information mining device different tests were performed.74.8698% and 76.9565% precision is given by utilizing Cross Validation and Percentage Split separately by J48. Naive Bayes gives 79.5652% exactness by utilizing PS. Algorithms gives most elevated accuracy by utilizing rate split test.

The researchers Sarwar and Sharma [9] (Sarwar.A, 2012) have played out a work on Naive Bayes for foreseeing diabetes Type-2.There are 3 sorts in Diabetes disease.Type-1 diabetes is the principal sort, Second sort is Type-2 diabetes, gestational diabetes is the third sort. Sort 2 diabetes happens from the expansion of Insulin resistance. There are 415 cases in the informational set and for assortment purpose; information is gathered from different parts of society in India. For the improvement of model MATLAB with SQL server is utilized. 95% of exactness is accomplished by utilizing Naive Bayes.

A researcher Ephzibah [10] (Ephzibagh.E.P, 2011) has built up a model for analysis of diabetes. This model is the mix of the GA and fuzzy logic. It is utilized to choose the best subset of features and furthermore utilized for the ssupgrade of exactness in classification. The dataset is taken from UCI Machine learning research facility which has 8 qualities and 769 cases for test. For implementation MATLAB is utilized. Just three best highlights/characteristics are chosen by utilizing genetic algorithm. Fuzzy logic classifier utilizes these three qualities and give 87% rightness.

The researchers Meherwar Fatima and Maruf Pasha [11] (Meherwar Fatima, 2017) in 2017 have done work on how machine learning is so essential in disease determination and its precision for expectation of illnesses in which pattern recognition is learnt from cases. The model is utilized for decision making process in anticipating the disease. This paper gave the investigation between diseases like coronary illness, diabetes sickness, liver infection, dengue malady and hepatitis ailment. At last they have inferred that statistical models are unsuccessful to hold the categorical information which assumes critical part in sickness expectation.

A researcher Anju jain[12] (Jian, 2015) in 2015 has done work on how extraction of data from various sources contain problems like heterogeneous data which is unorganized and high dimensions which can have missing data and outliers. Mining the data accurately using data pre-processing techniques like feature scaling and such other techniques for noise removal and missing data and can be used to build the model which certainly improves accuracy of the model and it will be useful in more biological complex situations.

A researcher Berina Alic[13] (Alic, 2017) in 2017 has worked on comparative analysis of most commonly used disease prediction techniques that are Artificial Neural Networks(ANN) and Bayesian Networks(BN) on classification of diabetes in early stage. In which higher accuracy is achieved by Artificial Neural Networks (ANN) with 89.78% than compared with Bayesian Network (BN) 80.43% due to independent relation between observed nodes. So where ANN's are the best way for predicting the diseases.

The researchers Vijayarani and Dhayanand [14] (Vijayarani.S, 2015) have anticipated the liver illness utilizing Support vector machine (SVM) and Naive bayes Classification algorithms. ILPD dataset can be acquired by utilizing UCI. Informational set comprises of 560 occurrences and it additionally comprises of 10 attributes. Comparision is done in light of the precision and time taken for execution. Naive bayes gives 61.28% exactness inside 1670.00ms. 79.66% rightness is gotten inside 3210.00ms by utilizing SVM. MATLAB is utilized with the end goal of execution. SVM gives most elevated accuracy when contrasted with the Naive bayes for the expectation of liver disease. Regarding time taken for execution, Naives bayes takes less time when contrasted with the SVM.

A researcher Gulia et al. [15] (Gulia.A, 2014) has played out an examination on intelligent methods which are utilized to group the patients having the liver maladies . This examination has utilized an informational set which is taken from UCI. In this test WEKA which is an information mining device is used.It had likewise utilized another five savvy systems J48,Random Forest, MLP, SVM and Bayesian Network classifiers are

utilized. In the stage 1, all the picked algorithms are connected on the first informational set to acquire the level of accuarcy. In stage 2, a strategy called feature selection is utilized and connected on the whole informational set to get the subset of liver patients and all these picked algorithms are utilized for testing the subset of whole informational set. In stage 3 they have done the correlation of results before the feature selection and after the feature selection. After the feature selection, algorithms give the most astounding rightness as J48 gives the 70.669% of rightness, 70.8405% precision is acquired by utilizing MLP calculation, SVM gives 71.3551% accuracy, 71.8696% rightness is gotten from Random forest.Bayes Net offers 69.1252% rightness.

The researchers Manimeglai and Fathima [16] (Fathima.A.S, 2012) have done a work for the prediction of the sickness called Arbovirus-Dengue. The Data mining algorithms which are utilized by them are Support Vector Machine(SVM).Data set that is utilized for investigation is taken from the King Institute of Preventive Medicine which is of Chennai and overviews of numerous healing facilities and research facilities which is of Tirunelveli from India. It contains 5000 examples and 29 attributes.R venture form 2.12.2 is utilized for looking at the information. Accuracy that is gotten by SVM is 0.9042.

A researcher Karlik [17] (Karlik.B, 2011) gives an examination that demonstrates the comparision between back propagation classifiers and Naïve Bayes which is utilized for diagnosing hepatitis malady. There is a fundamental preferred standpoint in utilizing these classifiers. That is just little measure of information is utilized with the end goal of classification. There are a few sorts in hepatitis like A, B, C, D and E. These are caused by different infections. An open source programming called Rapid Miner has been utilized as a part of this examination. The informational set is gotten from UCI. Informational set comprises of 155 cases and 20 highlights. The quantity of attributes utilized as a part of this investigation are15.15.97% accuracy is acquired from Naive Bayes classifier.

A researcher Eystratios et al. [18] (Eystratios G.Keramidas, 2007) has given a USG image analysis method for the detection of boundary of thyroid nodule. Initially Region of Interest which is usually said as ROI has been selected. Thyroid Boundary Detection Algorithm which is known as TBD algorithm has been used. K-Nearest Neighbor (k-NN) algorithm has been selected as a most powerful and useful classification method. The works well on longitudinal USG images.

In 2015 a researcher P. Swathi baby [19] (Baby, 2015) has developed a model that takes kidney disease data set of patients and developed a model that can predict the type of kidney disease. The model used several classification algorithms like random forests, ADtrees,j48, K-means algorithms and compared the results upon statistics that showed Random forest gives better result than the other algorithms.

In 2017 a researcher Sumedh Sontakke[20] (Sumedh Sontakke, 2017) has developed to study and compare two methodologies like machine learning and Artificial Neural Networks (ANN) which they are compared on the deaths reported and classified based upon different types of liver diseases in which ANN got better results. The field of diseases diagnosis is going to have more number of advancements in the coming years.

In 2017 SK Razia [21] developed a framework model to diagnose the thyroid disease using machine learngin techniques. The unsuper vised learning and supervised learning are used to diagnose the thyroid disease and compared with the decision tree model ultimately the framework model is outperformed than the decision tree model.

3. Algorithms in ML for Disease Diagnosis

There are several algorithms used in two phases of disease prediction that are mainly classified into two phases

Pre-Processing: In pre-processing we have several techniques for cleaning of the data where we need to combine heterogeneous, high dimensional data which contains noise and missing data. In the next step we need to apply feature scaling for the data so that the new data when entered into the model can predict correctly. At the last step in this phase we need to apply dimensionality reduction techniques to combine the data

For reducing the dimensions we have several algorithms like

Principal component analysis (PCA):

In this method we use orthogonal transformation which is a statistical technique for converting possible set of correlated values known as principal components.

Linear Discriminant Analysis (LDA):

It is a strategy utilized as a part of insights for pattern recognition to discover linear combination of features in machine learning approach. The result of this will be used for linear classification.

In the next Phase contains several machine learning classification algorithms as

Decision Trees:

The decision tree is one of the most important and also most used classification algorithm. This algorithm

utilizes a divide and conquer strategy to build a tree. There are a set of occurrences which are related with a collection of attributes. A decision tree comprises of nodes and leaves in which nodes are tests on the estimations of a characteristics or attributes and leaves are the classes of an example that fulfills the given conditions. The outcome might be "true" or "false". Rules can be acquired from the way which begins from the root node and finishes at the leaf node and furthermore uses the nodes in transit as preconditions for the got rule, to foresee the class at the leaf. The tree pruning must be done to evacuate pointless preconditions and duplications.

K-Means:

It can be said as k-means grouping. This is a technique for Vector Quantization, which is initially from signal processing, which is known for cluster analysis in information mining. We can utilize the 1-closest neighbor classifier on the cluster centers that are acquired from k-means to group the new information into effectively existing groups.

KNN:

It can be called as K nearest neighbor. It is the clustering technique used for clustering of data. It can be considered as another version of K-Means. It doesn't use the mean and distance. Instead it is based upon voting of the nearest neighbors in the k-clusters.

4. Conclusion

The Machine Learning is a type of brute force mechanism which tries to find the correlation between the numerical attributes of inputs with matching outputs based upon the previous data. In other words there is no suitable algorithm that can be so good for using in disease prediction as there is more labeled data. So as of now there exists some limitations even for machine learning algorithms.

References

[1] Vikas Tiwari, T. (2Ol3). Design and implementation of an efficient relative model in cancer disease recognition". IJARCSSE.

[2] M.peleg, S.tu. (2OO6). Decision Support, Knowledge Representation and Management . IMIA.

[3] Khaleel, M. A. (2Ol3). A Survey of Data M ining Techniques on Medical Data for Finding frequent diseas es. IJARCSSE.

[4] Vembandasamy.K, S. D. (2Ol5). Heart Diseases Detection Using Naive Bayes Algorithm. IJISET, 44l-44.

[5] Chaurasia.V, P. (2Ol3). Data Mining Approach to Detect Heart Disease. . IJACSIT, 56-66.

[6] Parthiban.G, S. (2Ol2). Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. IJAIS, 25-3O.

[7] Tan.K.C, T. Y. (2OO9). A Hybrid Evolutionary Algorithm for Attribute Selection in Data Mining. . IJDKP, 86l6-863O.

[8] Iyer.A, S. (2Ol5).) Diagnosis of Diabetes Using Classification Mining Techniques. . IJDKP, l-l4.

[9] Sarwar.A, S. (2Ol2). Intelligent Naive Bayes Approach to Diagnose Diabetes Type-2. ICNICT, l4-l6.

[10] Ephzibagh.E.P. (2O11). Cost Effective Approach on Feature Selection using Genetic Algorithms and Fuzzy Logic for Diabetes Diagnosis. IJSC.

[11] Meherwar Fatima, M. P. (2Ol7). Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, l-l6.

[12] Jian, A. (2Ol5). Machine Learning Techniques For Medical Diagnosis. ICSTAT.

[13] Alic, B. (2Ol7). Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases. Mediterranean Conference On Embedded Computing.

[14] Pratuisha.K, Rajeswara Rao.D, A comprehensive study: On artificial-neural network techniques for estimation of coronary-artery disease,Journal of Advanced Research in Dynamical and Control SystemsVolume 9, Issue Special Issue 12, August2017, Pages 1673-1683.

[15] Gulia.A, V. R. (2Ol4). Liver Patient Classification Using Intelligent Techniques. IJCSIT, 5Ol1-5115.

[16] Fathima.A.S, M. (2Ol2). Predictive Analysis for the Arbovirus-Dengue using SVM Classification.. International Journal of Engineering and Technology, 52l-527.

[17] Vidyullatha.P., Rajeswara Rao D:Machine learning techniques on multidimensional curve fitting data based on r- square and chi-square methods,International Journal of Electrical and Computer EngineeringVolume 6, Issue 3, June 2016, Pages 974-979..

[18] Eystratios G.Keramidas, D. K. (2OO7). Efficient and effective image analysis for thyroid nodule detection. ICIAR, lO52-lO6O.

[19] Baby, P. (2Ol5). Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms . IJERT.

[20] Sumedh Sontakke, J. L. (2Ol7). Diagnosis of Liver Diseases using Machine Learning. ICEI.

[21] SHAIK.RAZIA, M.R.NARASINGARAO published "A Neuro computing frame work for thyroid disease diagnosis using machine learning techniques", Vol.95. No.9. Pages 1996-2005) ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195.

[22] P. Gopi Krishna, K. Sreenivasa Ravi "DESIGNING A MULTIPURPOSE RECONFIGURABLE WIRELESS NODE FOR BROADCASTING AND UNICASTING IN REREAL TIME APPLICATIONS" in International Journal of Pure and Applied Mathematics (IJPAM). Volume 115 No. 8 2017, 505-510.

[23] P Gopi Krishna, K Sreenivasa Ravi "IMPLEMENTATION OF MQTT PROTOCOL ON LOW RESOURCED EMBEDDED NETWORK" in International Journal of Pure and Applied Mathematics (IJPAM). Volume 116 No. 6 2017, 161-166.

[24] Dr. Seetaiah Kilaru, Hari Kishore K, Sravani T, Anvesh Chowdary L, Balaji T "Review and Analysis of Promising Technologies with Respect to fifth Generation Networks", 2014 First International Conference on Networks & Soft Computing, ISSN:978-1-4799-3486-7/14,pp.270-273,August2014.

[25] N.Prathima, K.Hari Kishore, "Design of a Low Power and High Performance Digital Multiplier Using a Novel 8T Adder", International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue.1, Jan-Feb., 2013.

[26] T. Padmapriya and V. Saminadan, "Improving Throughput for Downlink Multi user MIMO-LTE Advanced Networks using SINR approximation and Hierarchical CSI feedback", International Journal of Mobile Design Network and Innovation- Inderscience Publisher, ISSN : 1744-2850 vol. 6, no.1, pp. 14-23, May 2015.

[27] S.V.Manikanthan and K.srividhya "An Android based secure access control using ARM and cloud computing", Published in: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on 26-27 Feb. 2015,Publisher: IEEE,DOI: 10.1109/ECS.2015.7124833.

[28] M. Rajesh, Manikanthan, "ANNOYED REALM OUTLOOK TAXONOMY USING TWIN TRANSFER LEARNING", International Journal of Pure and Applied Mathematics, ISSN NO:1314-3395, Vol-116, No. 21, Oct 2017.

[29] K.Srikar ,M.Akhil ,V.Krishna reddy "Execution of Cloud Scheduling Algorithms" International Innovative Research Journal of Engineering and Technology ISSN NO: 2456-1983.Volume 2, Issue 4 June 2017.108-111.