Gagan Kumar et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 3, Issue 2, June 2016, pp. 128-132

A survey on Machine Learning Techniques in Health Care Industry

Gagan kumar[1], Rohit Kalra[2]

[1]Assistant Professor, Modern Institute of engineering & technology
[2]Research scholar, Modern Institute of engineering & technology

Abstract: Medical diagnosis is a complicated task and plays a vital role in saving human lives so it needs to be executed accurately and efficiently. An appropriate and accurate computer based automated decision support system is required to reduce cost for achieving clinical tests. This paper provides an insight into machine learning techniques used in diagnosing various diseases. Various data mining classifiers have been discussed which has emerged in recent years for efficient and effective disease diagnosis.

Keywords: Machine Learning, Data Mining, Disease Diagnosis, Classification

1. INTRODUCTION

These days, health care industry generates a large amount of complex data about patients regarding clinical examination, treatment report, hospital resource management records, electronic patient records, medicine etc which has become cumbersome to organize properly. Due to improper organization of the data, the quality of decision making is getting affected. This increase in data volume requires ways in which data can be extracted and processed efficiently. The accurate diagnosis of life-threatening diseases such as breast cancer, heart disease, liver disease etc is a very crucial task in medical science. The humans and computers can be integrated together to achieve best results for correct diagnosis of diseases by balancing the knowledge of human experts in related domains with the vast search potential of computers. This kind of difficulty could be resolved with the help of machine learning techniques. Computer based decision support system can play an important role in correct diagnosis and cost effective treatment.

The use of computers and information technology is being increasingly implemented in health care organization in order to help doctors in their day to day decision making activities. It helps doctors and physicians in diseases management, tests, medications and discovery of patterns and relationships among clinical and diagnosis data and as well as employ machine learning techniques.

This paper is organized as follows: Section 2 gives the Classification of Machine Learning techniques including supervised and unsupervised algorithms. Section 3 describes the work in the literature regarding classification algorithms for medical diagnosis of diseases. Section 4 concludes the survey.

II. CLASSIFICATION OF MACHINE LEARNING TECHNIQUES

Machine learning is a domain of artificial intelligence involving the construction of algorithms that automatically learns through experience and performance of algorithm gets improved with each experience [1]. Algorithm operates by detecting some pattern in input data and building a model based on input data to make precise predictions for new data. The machine learning techniques are based on identifying patterns from large data sets that provide support for predictions and decision making process for diagnosis and treatment planning. The machine learning techniques can be classified as follows:

- Supervised Learning Techniques: In case of supervised machine learning, algorithm induces a mapping function from given labeled training dataset to map new input data to its desired output [2]. Labeled training dataset comprises of examples, which is a pair of input data and its output value. The problems solved by supervised learning techniques are basically categorized as regression and classification problems. In a regression problem, input variables are mapped to continuous output function whereas in a classification problem, input variables are mapped to discrete categories. The supervised learning techniques are as follows:
  - Decision Trees: Decision tree is a tree-like structure where each internal node denotes a test on a predictive attribute and each branch denotes an attribute value. A leaf node represents predicted classes or class distributions. An unlabeled object is classified by starting at the topmost (root) node of the tree, then traversing the tree, based on the values of the predictive attributes in this object. Eg ID3 which uses information gain with statistical pre-pruning, C4.5 an advanced version of ID3, and the most popular decision-tree algorithm, ART which minimizes a cost-complexity function, See5 which builds several models and uses unequal misclassification costs and IFN (Info-Fuzzy Network) which utilizes information theory to minimize the number of predictive attributes in a decision-tree model. Information gain is the

difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top ranked features are chosen as the potential attributes used in the classifier.

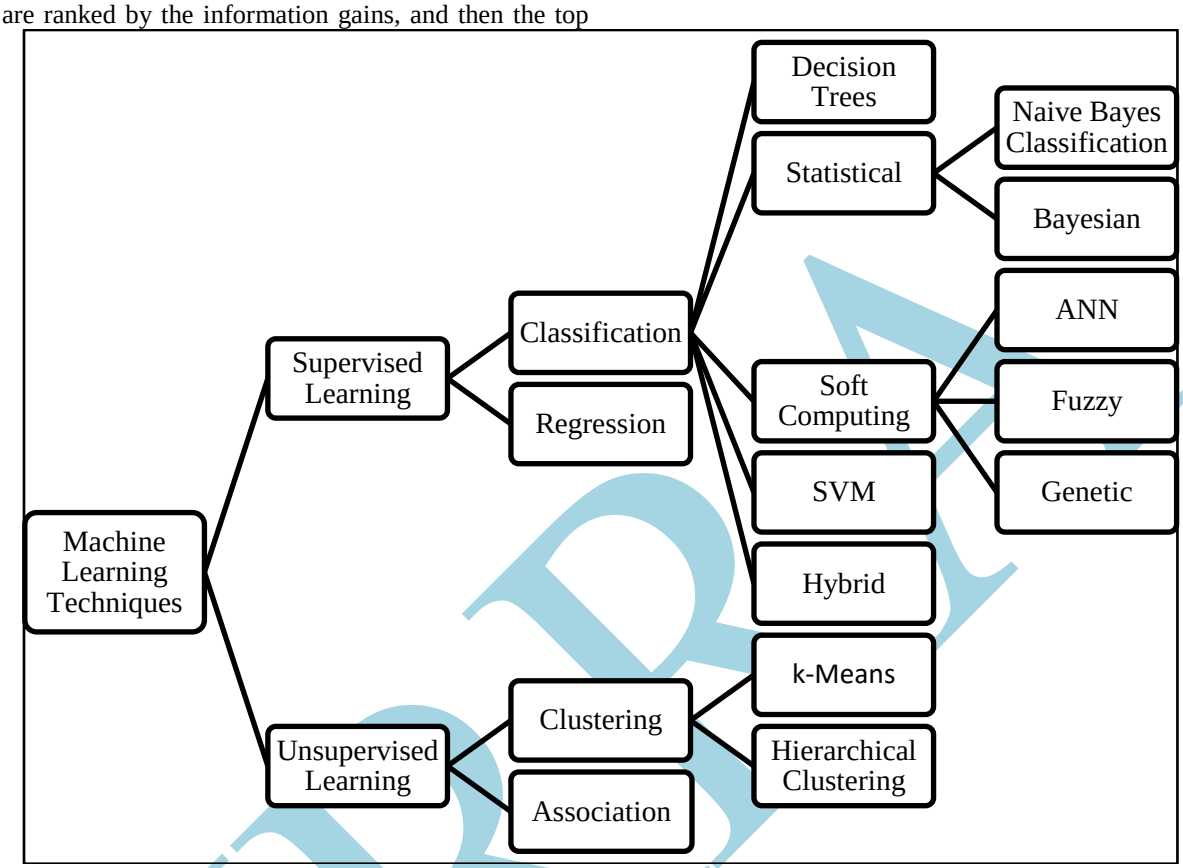

Figure 1. Classification of Machine learning techniques

- Soft computing Techniques: Soft Computing is the fusion of methodologies such as artificial neural networks (ANN), fuzzy algorithms and genetic algorithms [3/7] that were designed to model and enable solutions to real world problems, which are difficult to model mathematically. A neural network (NN) is a parallel, distributed information processing structure consisting of multiple numbers of processing elements called node, they are interconnected via unidirectional signal channels called connections. Each processing element has a single output connection that branches into many connections; each carries the same signal i.e. the processing element output signal. Eg Multilayer Perceptron Neural Network (MLPNN) and Back propagation algorithm network (BPN).
- Statistical Techniques: The statistical techniques are based on probability theory and provide a justification that a particular instance belongs to particular class with probability value P, $0 \leq P \leq 1$. Eg Bayesian networks and Naïve Bayes classifier [4].
- Support Vector Machines: Support Vector Machines (SVMs) are the newest supervised machine learning technique introduced by Vapnik in 1995. SVM simultaneously minimize the empirical classification error and maximize the geometric margin between the classes [5]. The RBF (Radial Basis Function) kernel of SVM can be used for classification of higher-dimensional data with less numerical difficulties [6]. Data sets are divided into test set and training set. Test data set is used to assess the performance of SVM classifier model created by training set. Validation using the test data set can avoid potential bias of the performance, estimate due to over-fitting of the model to training data sets.
- Hybrid techniques: The term hybrid means to combine two or more techniques or approaches to build a new approach.

- Unsupervised Learning Techniques: In case of unsupervised machine learning, algorithm infers a mapping function to find hidden patterns and correlation between them from unlabelled input dataset. Input dataset comprises of examples, each example is an input data with no explicit output value. We have very little idea of the output values in this case. We can find correlations by clustering the data as there is no feedback or teacher available for correction. For example, we have to discover close-knit group of friends in facebook. Eg

K-means clustering algorithm, hierarchical clustering [2].

III. RELATED WORK

Work done by various researchers in the field of disease diagnosis using data mining and machine learning techniques has been discussed below. Decision tree has been used as classifier for breast cancer diagnosis [7-9]. Azar and El-Metwally [10] proposed a decision support tool for the detection of breast cancer based on three different classifiers, namely, single decision tree (SDT), boosted decision tree (BDT) and decision tree forest. They claimed that BDT performed better than SDT with 98.83% and 97.07% accuracy respectively. The demerit of decision tree classifier in medical diagnosis is imbalancement and cost sensitivity problem.

Chowdhury et al. [11] utilized ANN in predicting neonatal disease diagnosis. The proposed technique comprised of Multi Layer Perceptron with a backpropagation learning algorithm for training ANN and recognizing a pattern for the diagnosing and prediction of neonatal diseases. The data set consists of 94 samples of different symptoms parameter. The technique exhibits ANN based prediction of neonatal disease with an accuracy of 75% with 64 training set and, 15 test set and 15 validation test.

Vanisree et al. [12], proposed a Decision Support System for diagnosis of Congenital Heart Disease. The proposed system is based on Backpropagation neural network which is multi layered Feed Forward Neural Network, which is trained by a supervised Delta Learning Rule. The dataset consists of 200 samples with 36 attributes each depicting signs, symptoms and the results of physical evaluation of a patient. The proposed system used 80% data set for training and 20% for testing and achieved an accuracy of 90% and mean square error of 0.016.

Ratnakar et al. [13] proposed a solution based on genetic algorithm for selection of optimal set of attributes for prediction of heart diseases and Naïve Bayes' technique to generate relationships amongst the attributes using the concepts of conditional probability. Using GA, 13 attributes are reduced to 6 which are further fed to Naïve Bayes Classifier. Chitra et al. [14] claimed that the application of ANN can be time-consuming due to the selection of input features for the multi layer perceptron. It is very slow training process and clinicians find it difficult to understand how its classification decisions are taken and cannot interpret the results easily.

Masethe [15] performed a comparison of various data mining algorithms on WEKA tool for the prediction of heart attacks to find the best method of prediction. The algorithms used are J48, REPTREE, Naïve Bayes, Bayes net and CART with prediction accuracy as 99.07%, 99.07%, 97.22%, 98.14%, 99.07% respectively. Archana and Sandeep [16] proposed a hybrid prediction model with missing value imputation (HPM-MI) based on K-means clustering with Multilayer Perceptron. The proposed algorithm was evaluated on three medical data sets, Pima Indians Diabetes, Wisconsin Breast Cancer, and Hepatitis from the UCI Repository of Machine Learning. The results show HPM-MI has produced accuracy, sensitivity, specificity, kappa and ROC as 99.82%, 100%, 99.74%, 0.996 and 1.0 respectively for Pima Indian Diabetes data set, 99.39%, 99.31%, 99.54%, 0.986, and 1.0 respectively for breast cancer data set and 99.08%, 100%, 96.55%, 0.978 and 0.99 respectively for Hepatitis data set.

Turabieh proposed a hybrid algorithm which integrates two powerful computational intelligence techniques namely Gray Wolf Optimization (GWO) and Artificial Neural Networks (ANN) for prediction of heart disease [17]. Gray wolf optimization is a global search method that works by minimizing the root mean square error while gradient-based back propagation method is a local search one. GWO is used for finding the initial optimal weights and biases for ANN model to reduce the probability of ANN getting stuck at local minima and slowly converging to global optimum. The performance of hybrid ANN-GWO is compared with normal ANN trained using back-propagation neural network. The results shown depict that the proposed model increases the convergence speed (time reduces to half) and the prediction accuracy. Tina Patil et al. [18] have applied two classification algorithm viz. Naïve Bayes based on probability and J48 based on decision trees to classify the item according to its features with respect to the predefined set of classes. The results demonstrate that J48 is more accurate and cost efficient than Naive Bayes algorithm.

Sunila et al. [19] proposed an improved Multilayer perceptron algorithm (MLP) which works on multiple subsets of training set. The majority probability rule is used to combine the results from different subsets. The experiment is implemented with 10-fold cross validation on Cleveland, Switzerland and Hungarian datasets. The result shows that proposed approach is better than MLP algorithm and has attained an accuracy of 82.8%.

Zheng et al [20] proposed K-means algorithm to recognize the hidden patterns of the benign and malignant tumors separately. The membership of each tumor to these patterns is calculated and treated as a new feature in the training model. Then, a support vector machine (SVM) is used to obtain the new classifier to differentiate the incoming tumors. The proposed algorithm achieves the accuracy to 97.38% with 10-fold cross validation.

Vadicherla et al. [21] suggested a sequential minimal optimization (SMO) technique of SVM for heart disease diagnosis system. The system is proposed for two classes. SMO helps in training of SVM by finding the optimal values of multipliers required during training phase. The result reveals that SMO shows good results even on large dataset and performance time is also improved. Kumari et al. [22] compared various data mining classification techniques RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) on cardiovascular disease dataset. The performance of above techniques is compared through sensitivity, specificity,

accuracy, error rate, True Positive Rate and False Positive Rate. The results obtained reveals error rates for RIPPER, Decision Tree, ANN and SVM as 2.756, 0.2755, 0.2248 and 0.1588 respectively, accuracy of RIPPER, Decision Tree, ANN and SVM as 81.08%, 79.05%, 80.06% and 84.12% respectively. So, it can be analyzed that SVM predicts cardiovascular disease with least error rate and highest accuracy.

IV. CONCLUSION

This survey provides the brief description of machine learning techniques for classification of diseases. The classification accuracy depends on the exact metrics which are used which also indicates the variety of features has been utilized. The role of classifier is important in healthcare industry so that the results can be used for determining the treatment. The existing techniques are studied and compared for finding the efficient and accurate systems.

V. REFERENCES

[1] Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. eds., 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[2] Celebi, M.E., Kingravi, H.A. and Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 40(1), pp.200-210.

[3] Tettamanzi, A.G. and Tomassini, M., 2013. Soft computing: integrating evolutionary, neural, and fuzzy systems. Springer Science & Business Media.

[4] N. Cruz-Ramírez, H. G. Acosta-Mesa, H. Carrillo-Calvet, L. Alonso Nava-Fernández, and R. E. Barrientos-Martínez, "Diagnosis of breast cancer using Bayesian networks: A case study," Comput. Biol. Med., vol. 37, no. 11, pp. 1553–1564, Nov. 2007.

[5] Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp.144 -152. ACM Press. 1992.

[6] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer Verlag. 1995.

[7] D. M. F. bin Othman and T. M. S. Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," in 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, F. Ibrahim, N. A. A. Osman, J. Usman, and N. A. Kadri, Eds. Springer Berlin Heidelberg, 2007, pp. 520–523.

[8] N. Cruz-Ramírez, H.-G. Acosta-Mesa, H. Carrillo-Calvet, and R.-E. Barrientos-Martínez, "Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks," Appl. Soft Comput., vol. 9, no. 4, pp. 1331–1342, Sep. 2009.

[9] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. C. Hsieh, "A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification," Appl. Soft Comput., vol. 11, no. 1, pp. 632–644, Jan. 2011.

[10] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013.

[11] D. R. Chowdhury, M. Chatterjee & R. K. Samanta, "An Artificial Neural Network Model for Neonatal Disease Diagnosis", International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 2, no. 3, pp. 96-106, 2011.

[12] Vanisree K, Jyothi Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications, vol. 19, no. 6, pp. 6-12, 2011.

[13] S. Ratnakar, K. Rajeswari & R. Jacob, "Prediction of Heart Disease Using Genetic Algorithm for Selection of Optimal Reduced Set of Attributes", International Journal of Advanced Computational Engineering and Networking, vol. 1, no.2, pp. 51-55, 2013.

[14] Anuja Kumari, V & Chitra, R 2013, 'Classification of Diabetes Disease Using Support Vector Machine', International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801.

[15] Masethe, H.D., Masethe, M.A.: Prediction of heart disease using classification algorithms. In: World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, San Francisco, USA, 22–24 Oct 2014.

[16] Purwar, A. and Singh, S.K. (2015) Hybrid Prediction Model with Missing Value Imputation for Medical Data. Expert Systems with Applications, 42, 5621-5631.

[17] Turabieh, H. , "A Hybrid ANN-GWO Algorithm for Prediction of Heart Disease", American Journal of Operations Research, 6, pp. 136-146, 2016.

[18] Tina Patil, R & Sherekar, SS 2013, 'Performance Analysis of Naive bayes and J48 Classification Algorithm for Data Classification', International Journal of Computer Science and Applications, vol. 6, no.2, pp. 256-261.

[19] P. Panday and N. Godara, "Decision Support System for Cardiovascular Heart Disease Diagnosis using Improved Multilayer Perceptron," International Journal of Computer Applications ,vol. 45, no. 8, pp. 12–20, 2012.

[20] Zheng, B., Yoon, S.W. and Lam, S.S., 2014. Breast cancer diagnosis based on feature extraction using a

hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), pp.1476-1482.

[21] E. Technologies, D. Vadicherla, and S. Sonawane, "Decision Support System for Heart Disease Based on Sequential Minimal Optimization in Support," International Journal of Engineering Sciences and Emerging Technologies, vol. 4, no. 2, pp. 19–26, 2013.

[22] Ishtake, S.H., Sanap, S.A.: Intelligent heart disease prediction system using data mining techniques. Int. J. Healthc. Biomed. Res. 1(3), 94–101 (2013)

