IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Issue-3 E-ISSN: 2347-2693

Analysing the supervised learning methods for prediction of healthcare data in cloud environment: A Survey

N.M. Annigeri [1], S.Shetty [2], A.P.Patil[3*]

[1 2 3*] Dept. of CSE, Ramaiah Institute of Technology, Bengaluru-54, Karnataka, India

[*]*Corresponding Author: nagashree.annigeri@gmail.com, Tel.: +91-9844727051*

Available online at: www.ijcseonline.org

Received: 23/Feb//2018, Revised: 28/Feb2018, Accepted: 21/Mar/2018, Published: 30/Mar/2018

Abstract— In the present era of massive usage of computers, an enormous set of data is being generated from different organizations each day, each hour and each second. This data would be of prodigious use to a diverse set of people based on their needs. Predictive analysis is a process of analysing data and identifying the different patterns in it, so as to predict the occurrence of these patterns in future. The predicted output can help plan a new strategy and adopt innovative solutions for the decision making. This paper attempts to analyse the various predictive models which are applied in the healthcare domain. These models are analysed in depth and will be proposed to be available on the cloud environment in future and can be accessed by those concerned for potential analysis.

Keywords— Predictive modeling, predictive algorithms, predictive analytics in a cloud environment, supervised learning.

I. INTRODUCTION

Healthcare domain has been one of the most interesting areas for researchers to experiment with predictive analytical techniques. With the advent of cloud computing infrastructure, the data which is generated at different health care units can be made available on a distributed environment.

The inherently large size of data generated by several healthcare organizations is throwing numerous open challenges for researchers. They are forced to think of newer and better techniques for managing and organizing the humongous data which is generated every second.

Research findings in the area of predictive analytics over cloud environment would provide great insights to healthcare firms to overcome the problems associated with traditional database architecture. Presently many healthcare firms use the traditional database systems to store and maintain the doctor's and patient's data. But as and when the data collected increases over time due to growing population and their health issues, the space required to store this data is more and adds up the cost of hardware and maintenance. It needs substantial amount of time for the data analyst to split the existing data. However literature states that about 80% of the available data is a suitable data and can be used to future analysis. 20% of the data would not be suitable for analysis as it may contain unfitting or junk data, which further requires to be studied for its relevance. With the aid of predictive analytical techniques classification of this data can be done in lesser time. Predictive analytical solutions would provide decisive information which would help healthcare organizations to take right decisions at the right time.

Rest of the paper is organized as follows, Section II contain the related work/literature survey of the work being done in the domain of machine learning. Section III contains the general methodology followed. Section IV concludes survey work with future directions.

II. RELATED WORK

A number of research literature is available in the healthcare domain and the usage of predictive models for the same. Many such models like regression models, classification models and few other aspects are presented in this paper.

A. Predictive Data Analytics

Predictive data analytics is defined as the collection of machine learning methods/techniques as well as the statistical methods applied on the datasets to predict the future [1]. Machine Learning techniques help in a better way to gain all the useful information from the given data so that corrective actions can be taken to improve revenue, quality, and customer satisfaction [2] in general.

B. Predictive Data Analytics in Healthcare

Data Analytics is finding its extensive application in healthcare industry. It attempts to provide proof based medication which aids in the appropriate diagnosis of the diseases and their associated treatment. In healthcare domain, predictors are used to anticipate the effectiveness of the treatment.

The medical industry has evolved with a tremendous amount of data which are saved in the form of papers to a much more digital one. As the world is getting digitized, papers are getting replaced by e-copies, e-forms etc. Some of the advantages of applying analytics to this data in the healthcare industry can be listed as below:

- It would be easy to detect diseases based on specific patterns and these diseases can be given treatment early to ensure its effectiveness.
- The advancement or developments of many of these diseases can be predicted and can be addressed with analytical techniques.
- It helps in achieving patient churn and be able to provide leaner, much faster treatments to the patients.
- By adopting the predictive models in determining the diseases, there will be less room for human errors during treatment of patients.
- By analyzing the patterns of diseases, the doctors and hospitals will be able to provide better health observation and better responses to the treatments.
- It also helps in giving proof based medicines.
- The data collected with respect to the diseases; the necessary treatment information; planning of the future treatment-will all be available on a distributed cloud environment. By this the geographic barriers can be totally avoided. This would ease the work of the both hospital administration and patients to a great extent.

Paper [3] describes how predictive analytics can help the researchers to gain novel and deep insights to clinical and organizational decision making. The paper focusses on different learning algorithms such as classification, clustering and association and lists out major benefits and limitations of each algorithm. The authors discuss the algorithms such as Naïve Bayes, Artificial_Neural_Network, SVM, Decision Tree, Ensemble, and AdaBoost under Classification. Apriori algorithm under Association. The authors also mention about guidelines followed while performing the execution of the above algorithms. The author concludes that the decision tree performs well among classification algorithms because of its ability to visually represent the classification decisions made.

Thanh Nguyen *et al* [4], proposed an algorithm called fuzzy-standard-additive-model (f-SAM) along with the incorporation of Genetic algorithm (GA) to address ambiguity, computational difficulties faced by high-dimensional medical data collected. The hybrid integration technique GSAM is implemented against Probabilistic neural network, Adaptive vector quantization (AVQ), Fuzzy ARTMAP. Their implementation results show the dominance of the GSAM method over other models in terms of accuracy, F-measure and AUC (area under characteristic curve) [4].

The authors in the paper [5], discuss the methods of machine learning for analyzing tumor features in breast screening. The data is mined after dynamic_contrast_enhanced/magnetic_resonance_imaging (DCE-MRI). They demonstrated the algorithms such as k-means, self-organizing-maps (SOM) in order to examine the signal structures. Decision trees, support vector machines and K-nearest-neighbor (k-nn) are used under classification. Their result showed that self-organizing-maps performed well followed by k-nn and decision trees with respect to smoothness energy function for projection quality and classification accuracy in sensitivity and specificity.

Shipeng Yu *et al* [6], in their experiment applied a novel framework on different hospital data to evaluate the patient readmission risk. They considered the classification models support vector machines and cox regression along with LACE model (industry standard method which is followed by hospitals). LACE model is described as: L-duration (length) of visit, A-severe ness (acuity) of admission, C-comorbidity in index list, E-total visits in case of emergency during last few months. In this work, the framework designed is more effective and flexible compared to LACE model with respect to hospital readmission risk prediction [6].

The objective of the paper [7] is to predict heart related complications based on the existing patient history. The authors discuss the machine learning techniques namely support vector machines, AdaBoost, logistic regression, k-likelihood ratio test and Naïve Bayes. They evaluate their models with respect to prediction accuracy and AdaBoost performed to be the best among all techniques under prediction accuracy followed by support vector machines.

The authors Guilia Toti *et al* [8], in their work demonstrates the association rule based technique Apriori algorithm in R environment to analyze the correlation between asthma in infants and exposure to pollutant mixtures. The authors collect the data related to asthma infected infants and

associated them with six pollutant levels recorded earlier. The result showed that Apriori algorithm described 27 rules that reduced the false detection rate (FDR). FDR is proved to be less than 13%. The result also showed that the provision of the guidelines differs from 0.54% to 5.82%.

In the paper [9], the authors Dursun Delen *et al*, Glenn Walker, Amit Kadam experimented with the algorithms artificial neural networks, decision trees along with logistic regression to develop the predictive model on the breast cancer dataset. The authors also adopted cross-validation techniques (10-fold) in order to measure the performance. Metrics considered for the performance measures are accuracy, sensitivity and specificity. The results in their experiments show that the decision tree performed with highest accurateness of 93.6% followed by ANN with the correctness of 91.2% and lastly logistic regression with the lowest accuracy among three 89.2%.

Luis Claudio de Oliveira Silva *et al* [10], in their paper discuss the experiments conducted to detect abrasions in dense breasts and for the analysis the authors take two public databases with two groups of regions of interest(dense breasts and non-dense breasts). Independent component analysis and principal component analysis techniques are used as part of experiments. The paper concludes that independent component analysis performed better with an accuracy of 92.71% in dense breasts and 79.17% in non-dense breasts when compared to principal component analysis in detecting abrasions in the dense breasts.

C. *Predictive Data Analytics in Cloud environment*

With advancement of cloud and mobile expertise, data has advanced immensely with respect to quality and volume in order to fulfill a diversity of user requirements. Handling as well as treating such huge data proficiently and effectually becomes a chief task for any industry. The probability of pay-per-use with on-demand operations by cloud service providers is gaining fame in readiness computing model.

A considerable amount of work has been done in the field of cloud computing in few years.

Haluk Demirkan *et al* [11], in their paper discuss about a conceptual-framework for decision support system and also the fundamental foundation elements for service-orientation in cloud environment. The authors provide major requirements for a decision support system to implement and information about the unique model called Analytics as a service. The paper does not discuss the impact of the service-orientation on the working of decision support system in cloud.

In the paper [12], the authors Karsten Molka and James Byrne discuss the predictive cost models for cloud ecosystem. Multi~cloud of the cloud_federation, Private_ cloud, Cloud_bursting and cloud_brokerage were considered for the cost analysis. The authors work on the parameters like real-time investigation of the facility, virtual info and physical info of the source from the cloud setup in order to carry out the economic and predictive models evaluation which can forecast upcoming facility financial tendencies and consequence on entire price of proprietorship at infrastructure supplier side.

The authors Bo Jin *et al* [13], in their study propose a trust model based on the combination of the cloud model and Bayesian network. They evaluate the trust model using naïve-Bayesian network where conditional probability tables are drawn for the values. An algorithm is designed for the context free trust evaluation. With the help of the algorithm and experimental results, the authors suggest that anyone can use the Bayesian network to build the trust model and it helps in selecting the service provider.

The authors Deval Bhamare *et al* [14], makes an analysis on the feasibility of the supervised learning for the security of the cloud ecosystem. Authors use the datasets UNSW and ISOT on which the performance of the machine learning models is evaluated. The algorithms include decision trees, regression models, naïve-Bayes and support-vector-machines. The authors stress on the fact that the machine learning models which are tested with particular dataset have to be tested with other datasets also to check the robustness. The authors determine in their experiment that logistic regression performed best with an accuracy of 89.26% succeeded by J48 with 88.67%.

The authors Weider D. Yu *et al* [15], in their experimental study, propose a model of the distributed-storage-solution for hybrid cloud designed to load-store-retrieve the voluminous digital data in an effective way for a cloud based healthcare system. The authors largely focus on blending the approaches of RDBMS and NOSQL instead of using the solitary approach. The proposed system consists of a huge number of all the combination of structured, unstructured and semi_structured data. In this process, private cloud is used to collect critical data related to patients' as well as the data related to the public is stored in public-cloud. The author analyzed and observed the time taken in loading and retrieving the data from Amazon storage and Network Attached Storage (NAS). The author concludes

that Amazon is faster because of its powerful virtual instances and NAS was equally competitive due to its ease of operation, storage capacity and security.

The authors Kiran Rao and Sandeep Kumar [16] in their paper explores and discusses about the application of machine learning methodologies on resource-monitoring, resource-provisioning and management of cloud ecosystem. The machine learning approaches such as linear regression and reinforcement learning are used for virtual machine mapping and resource provisioning respectively to develop a framework for self-resource monitoring in cloud environment.

III. METHODOLOGY

The methodology or the procedure to be followed is as follows:

Step1:
Define the dataset: The dataset is created which satisfies the objectives of the project. This dataset describes all the necessary information related to the specific domain. Convert the dataset into digital format using Microsoft excel.

Step2:
Data preprocessing: After the collection of the dataset, apply preprocessing techniques (ex: feature selection, correlation – if applicable) and make the data ready for algorithm implementation.

Step3:
Test the model: Apply the supervised learning models on the preprocessed data.

Step4:
Evaluation of the output: Learning models are compared according to the performance metrics such as accuracy, error rate, ROC.

Step5:
Store and Retrieve from cloud: All the predictive models / datasets developed are stored in the cloud ecosystem. These details can be accesses from the cloud as per the user's needs.

The main advantage of supervised learning methodology with cloud access is to get cloud based ready solution for the user according to their individual requirement. The user can access the predictive models on "on-demand" basis. It helps the user or the organizations to take rapid decision by reducing the excess man-hours thus saving time, energy and increases the productivity.

Cloud access leverages the usage of facilities which the cloud service providers offer.

IV. CONCLUSION AND FUTURE SCOPE

This paper mainly contemplates on the different machine learning methodologies available for use in the field of healthcare. Supervised learning algorithms can perform tasks based on the predefined set of rules provided by the user. The review on the available literature indicates that the results obtained by applying these predictive techniques on the healthcare data sets will provide insights on the basic attributes and parameters of the same. These predictive models will further play a very important role in recommending the most suitable machine learning technique for the given dataset.

Here an attempt is made to give an exhaustive analysis of the application of machine learning algorithms to several datasets pertaining to healthcare domain. Finally we have summarized the advantages and feasibility of deploying the predictive models and the data sets on the cloud environment [17]. This indicates a futuristic use of the same to analyst and end-users.

REFERENCES

[1] J. D. Kelleher, B. M. Namee and A. D. Acry, "*Fundamentals of machine learning for predictive analytics: algorithms, worked examples and case studies,"* MIT Press, 2015.

[2] Siegel, Eric Author. "*Predictive Analytics: the Power to Predict Who Will Click, Buy, Lie or Die."* Wiley.

[3] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. and Hua, L. (2011). *"Data Mining in Healthcare and Biomedicine: A Survey of the Literature.* "Journal of Medical Systems, 36(4), pp.2431-2448.

[4] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, *"Classification of healthcare data using genetic fuzzy logic system and wavelets," Expert Systems with Applications,* Elsevier publicaitons, vol. 42, no. 4, pp. 2184–2197, 2015.

[5] T. Nattkemper, B. Arnrich, O. Lichte, W. Timm, A. Degenhard, L. Pointon, C. Hayes and M. Leach, *"Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods",* Artificial Intelligence in Medicine, Elsevier publication, vol. 34, no. 2, pp. 129-139, 2005.

[6] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand and B. Krishnapuram, *"Predicting readmission risk with institution-specific prediction models",* Artificial Intelligence in Medicine, Elsevier publications, vol. 65, no. 2, pp. 89-96, 2015.

[7] W. Dai, T. Brisimi, W. Adams, T. Mela, V. Saligrama and I. Paschalidis*, "Prediction of hospitalization due to heart diseases by supervised learning methods",* International Journal of Medical Informatics, vol. 84, no. 3, pp. 189-197, 2015.

[8] G. Toti, R. Vilalta, P. Lindner, B. Lefer, C. Macias and D. Price, *"Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining",* Artificial Intelligence in Medicine, Elsevier publication, vol. 74, pp. 44-52, 2016.

[9] D. Delen, G. Walker and A. Kadam, *"Predicting breast cancer survivability: a comparison of three data mining methods",* Artificial Intelligence in Medicine, vol. 34, no. 2, pp. 113-127, 2005.

[10] L. de Oliveira Silva, A. Barros and M. Lopes*, "Detecting masses in dense breast using independent component analysis",* Artificial Intelligence in Medicine, vol. 80, pp. 29-38, 2017.

[11] H. Demirkan and D. Delen, *"Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud",* Decision Support Systems, Elsevier B.V publications, vol. 55, no. 1, pp. 412-421, 2013.

[12] K. Molka and Byrne, James, *"Towards Predictive Cost Models for Cloud Ecosystems",* poster paper, IEEE Research challenges in Information Science, Paris, 2013.

[13] B. Jin, Y. Wang, Z. Liu and J. Xue, *"A Trust Model Based on Cloud Model and Bayesian Networks",* Procedia Environmental Sciences, vol. 11, pp. 452-459, 2011.

[14] Anitha H M, P. Jayarekha , "*Security Challenges of Virtualization in Cloud Environment*", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.1, pp.37-43, 2018.

[15] Weider D. Yu, Manjula Kollipara, Roopa Penmetsa, Sumalatha Elliadka, *"A distributed storage solution for cloud based e-Healthcare Information System",* 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013),Lisbon,Portugal 2014.

[16] P Kiran Rao, R Sandeep Kumar*, "Machine Learning Methods for Cloud Computing",* i-manager's Journal of Cloud Computing , Vol. 3 No. 4 August - October 2016 .

[17] K.Sree Divya, P.Bhargavi, S.Jyothi *"Machine Learning Algorithms in Big data Analytics",* International Journal of Computer Sciences and Engineering, vol.6, issue 1, 2018.

Authors Profile

Nagashree M Annigeri completed her B.E in computer science from Kuvempu University, Karnataka in 2006. She is currently persuing her MTech at Ramaiah Institute of Technology, Bengaluru, Karnataka in Computer Science and Engineering. Her areas of interest include Software Engineering, IoT, Data Analytics, Operating systems.

Savita K.Shetty received her B.E. (1996) in Computer Science and Engineering from Karnatak University ,Dharwad, and M.Tech (2004) in Computer Science and Engineering from Visvesvaraya Technological University,Belagavi. She is currently working on her Ph.D. in Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, Karnataka. Her research interests include Data Analytics, Data Mining and Machine learning.

Annapurna P Patil received her B.E from Gulbarga University, Karnataka in 1994, M.E and PhD from Visvesvaraya Technological University, Karnataka in 2001 and 2014 respectively. She is currently Professor in Computer Science Department of Ramaiah Institute of Technology. She is also a Senior IEEE member, ACM member, LMISTE, LMCSI. Her areas of interest include Wireless Networks, Data Analytics, Artificial Intelligence, Protocol Engineering, Cloud Computing and IoT, Data Structures and Analysis of Algorithms.

