EasyChair Preprint

№ 3116

Multi-Disease Prediction and Classification Based on Medical History

Abhishek Gupta and Kushagra Gupta

April 5, 2020

MULTI-DISEASE PREDICTION AND CLASSIFICATION BASED ON MEDICAL HISTORY

ABHISHEK GUPTA

B. Tech, Computer Science and Engineering. SRM Institute of Science and Technology, Kattankulathur
abhishekgupta98@gmail.com

KUSHAGRA GUPTA

B. Tech, Computer Science and Engineering. SRM Institute of Science and Technology, Kattankulathur
kushagragupta0802@gmail.com

ABSTRACT: This project intents to identify a person's risk of different Chronic Disease based on their Medical History and Genetic Predispositions. In this the user is able to enter their own medical records and family history and on the basis of such with help of Classification method on the basis on medical data we are able to predict presence of diseases. The project contains major diseases such as Chronic Kidney Disease, Cardiovascular Disease, Liver Disease. By providing an easy detection it prevents further worsening of the disease and their side effects if caught in earlier stages. It will be built using KNN and Inverse frequency Algorithm. The entire framework will be deployed on an online framework build on HTML5 and hosted on a Live Server in real time.

KEYWORDS: Chronic Kidney Disease (CKD), Prediction, Classification, Data set.

INTRODUCTION

Medical care is one of the most important aspect of the human society. We are all susceptible to multiple disease which cause physical ailments and biological deterioration. According to a survey by the Physicians Foundation, doctors on average see 20 patients a day.

Health care should be more about proactively identifying a disease and its risks then reacting to them. In the project we plan to use the K- NN (Nearest Neighbors) Algorithm which gives high accuracy for prediction of disease without the need for physically invasive tests.

The objective of the project is to analyses and detect the possibility of a disease in a person on the basis of indicators. It will detect the possibility of diseases with higher accuracy then possible otherwise without invasive lab tests. The final ruling will have to be done with proper lab testing. The proposed system will increase the probability of finding the disease and thus reduce unnecessary invasive tests done on the patient.

We take the different indicators of Chronic Diseases like the BP, Kidney function test, Complete blood count (CBC), Blood Sugar, ECG, Sera Creatinine etc. from the available data set. The data set is cleaned and missing values are estimated using root mean square function. We use linear regression models in python to model the data. We run a classification algorithm on the data sets to train the model with the given data. Supervised machine learning algorithms like KNN are used to train the model. The output will be a percentage of accuracy of correctly predicted cases of diseases in the test set. We expect above 95% accuracy in the sample test data and up to 90% in wider actual application.

We attempt to minimize errors by using cross validation method to prevent over fitting of data.

RELATED WORK

Chronic Diseases are very complicated and long lasting and can be managed with focused treatment easily if caught in early stages. The main objective of this survey is to predict the presence of a Chronic Disease in its early stages without the need for multiple invasive tests. In order to achieve that systems that use CNN or different machine learning algorithms are proposed which use different classifiers to analyses and predict diseases are applied. They are then graphed, and the accuracy is calculated to find the most efficient method available. They use conventional algorithms without and bias correction which reduces the true accuracy of the model.

LIMITATIONS OF THE PRESENT SYSTEM

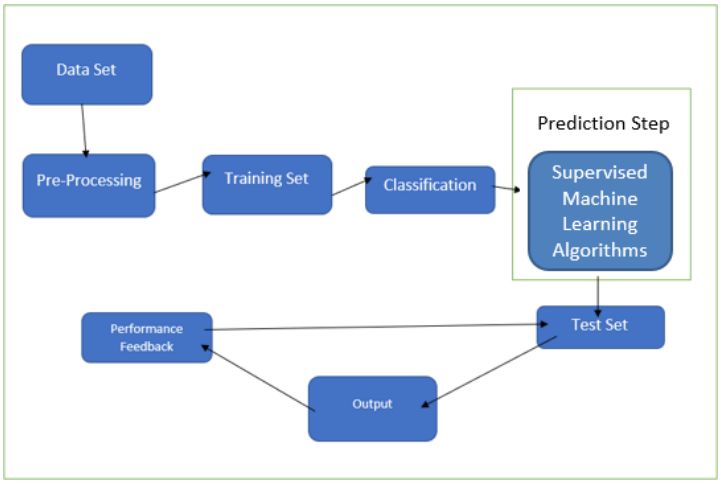
The present system includes a lot of unrelated factors which increase error in detection causing a bias in the data which reduces the accuracy of their model. This is because, while using KNN as a classifier if unwanted or useless data is used overfitting or underfitting occurs. However, sometimes it helps in providing more accurate predictions from seemingly useless indicators. Therefore, in certain circumstances it is helpful but may be a side effect of small dataset availability in medical field due to legal reasons. The present system is seemingly limited because of its unavailability of data and lack of checking algorithms to prevent over or under fitting. It does not talk about the relative accuracy of similar prediction models. Therefore, it is safe to say that there may exist a more accurate outcome.

BENEFITS OF THE PROPOSED SYSTEM

The project will implement checking and bias prevention methods such as cross-validation on the same dataset, trying to find the difference in accuracy of the model the was previously caused due to a bias. The most accurate method will then be implemented with newly collected data. It will also use seemingly unrelated parameters into consideration which are not linearly correlated and expanding the scope of the system and eliminating bias.

The model will also expand on the range of possibilities of the system by running it over new dataset and predicting corelated diseases. Numerical output is expected of the model with the related accuracy. The model can be incorporated in a Graphical User Interface to be easily usable by laymen to predict disease on the biases of their medical tests and health.

ARCHITECTURE DIAGRAM



LITERATURE SURVEYS

RESEARCH PAPER 1- BASE PAPER

PAPER NAME: Chronic Diseases Prediction over Bigdata by using Machine Learning

Author: Shreekanth Jogar, Pavankumar Naik (2019)

ABOUT THE PAPER: With the growth of big data in healthcare and healthcare communities comes accurate analysis of medical data which facilitates early diagnosis, easy patient care and other medical predictive services. However, the accuracy of the analysis is reduced when the quality of medical data is still incomplete. In addition, different regions show different characteristics of specific regional diseases, which may affect the prognosis of disease outbreaks. In the paper, the authors have used a supervised algorithm for successful diagnosis of chronic diseases in general disease populations. They tested modified models of real-time hospital data collected from central China in 2013- 2015. To

overcome the limitations of incomplete data, they used a latent factor model to reconstruct missing data. At the time of this paper there is no existing work focusing on analyzing factors to predict a chronic disease. In comparison to the some of the genera prediction algorithms, the prediction accuracy of their proposed algorithm gives 94.8% accuracy and is faster than CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

RESEARCH PAPER 2

PAPER NAME: Predicting individual disease risk based on medical history

Author: Darcy A. Davis, Nitesh V. Chawla (2008)

ABOUT THE PAPER: The cost of health care specifically for treatment of chronic illnesses keeps rising to the point of becoming affordable. The crisis has created a drive for more preventative medicine model, when the primary concern is understanding risk of a disease and taking early actions. However, comprehensive testing is very costly and consumes a lot of precious time of both the patients and the doctor. In this paper they have proposed CARE which is an Recommendation Engine it depends only on the patient's medical and genetic history and using ICD-9-CM codes to present the predicted future chronic diseases risks. Their system uses collaborative filtering to define a patient's highest risks based on their medical and family genetic history and those of patients with similar history. The used an Iterative model named ICARE, which uses the results of a past prediction to improve future accuracy and performance. Even though these systems require no specialized structured information for predicting medical and genetic conditions on a ranking basis. They present results as a structured dataset, showing that CARE and ICARE perform well at storing future disease risks. They rake these data mining and big data algorithms to predict future patients and their highest risks to improve the condition of preventive care. Even though their work, can lead to new standards in predictive care it is held back due to lack of data and the different privacy polices present in the medical industry causing a bottleneck in acquiring the initial data required to get the system started with a sufficiently high accuracy.

RESEARCH PAPER 3

PAPER NAME: An analysis on Chronic Kidney Disease prediction system: cleaning, pre- processing and effective classification of data.

Author: Ankit, Bhagyashree Besra and Banshidhar Majhi

ABOUT THE PAPER: Chronic Kidney Disease or CKD is a condition in which the kidney stops functioning normally which causes a lot of other harmful symptoms for the patients. Like other chronic diseases CKD also benefits with early detection. A prolonged duration of undetected CKD may deteriorate the kidney to the point of kidney failure. Some of the major symptoms of CKD are anemia, nerve damage, high blood pressure, weak bones, heart, etc. In this paper, the authors have proposed a system that can be used to predict CKD with a very high accuracy and followed by the prediction of kidney damage percentage. Their main objective is to predict the disease early on for easier management of the symptoms and also

reducing the need for a lot unnecessary false testing. They starts their perdition with data pre-processing and ends with the disease classification as well as identifying the classified result. The result estimates presence as well as the stage of CKD so that the treatment of the patient can be done accordingly.

RESEARCH PAPER 4

PAPER NAME: Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data.

Author: Ravizza, S., Huschto, T.

ABOUT THE PAPER: In this paper the authors have used real world data for the diagnoses process, therapeutic suggestions, and risk predictions for diabetes and blood pressure based chronic kidney disease on strictly controlled clinical trials. They show a study by study comparison of prediction of theses disease using their real world-based system and how it beats published algorithms on clinical study data.

RESEARCH PAPER 5

PAPER NAME: Smart collaboration framework for managing chronic disease using recommender system

Author: Asmaa S. Hussein, Xue Li Katarina Ivani´c Mladen Jardas

ABOUT THE PAPER E-Healthcare services show high-quality solutions to reduce patients' health risks by improving the services provided and by having cost effective solutions. This paper presents a Chronic Disease Counselor and a prognostic system that helps patients to monitor and control their cases by demonstrating prognosis and medical solutions. In order for the system to be accurate, which is important for that model, it needs to work with high quality data. This work proposes an integrated filtering framework that reduces complexity and time taken and improves response time for high quality data to improve the client forecasting process. This paper concludes by presenting a real clinical study with real medical data on providing medical advice and diabetes solutions.

RESEARCH PAPER 6

PAPER NAME: Empirical analysis of predictive algorithms for collaborative filtering. Technical Report MSR-TR-98-12, Microsoft Research, May 1998

Author: J.S. Breese, D. Heckerman, and C. Kadie.

ABOUT THE PAPER: Collaborative filtering and prediction systems use a database about customer preferences to predict additional medical topics or useful products that a user may like. In this paper we describe several algorithms developed for this work, including techniques based on combinations of coefficients, calculation of vector similarity, and Bayesian mathematical methods. We test the theoretical accuracy of the various methods in the domain of representative problem fields. We use two basic metric test classes. This metric

uses the probability that the client may have seen a prediction on a pre-determined list. It was used to test the data associated with the three application areas, the four test principles, and the two-test metrics for the multiple algorithms. The results indicate that in different contexts, between interacting with Bayesian networks, the direction of choice depends on the nature of the data and the type of applications.

RESEARCH PAPER 7

PAPER NAME: "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 855– 864.

Author: N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka,

ABOUT THE PAPER: We propose a comprehensive construct based on multidisciplinary care (MTL) that focuses primarily on the prediction of patient mortality from clinical data. The MTL framework enables the model to read the user's presentation at various clinical prediction tasks. In addition, we show how MTL facilitates small but consistent benefits in a single phase using simply integrating related functions into the MTL framework. To find out, we use the multi-level Convolutional Neural Network (CNN) and the MTL loss component. The model is analyzed with three, five, and twenty functions and is able to produce a higher performance model than a single learning classifier. We then discuss the results of a multidisciplinary model on clinical outcomes of interest, including producing high-quality presentations that can be used to work effectively with simple models.

RESEARCH PAPER 8

PAPER NAME: Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network Periodicals of Engineering and Natural Sciences Vol.6, No.1, June 2018, pp. 229~240

Author: B Sankara Babu, A Suneetha, G Charles Babu, Y. Jeevan Nagendra Kumar, G Karuna (2018)

ABOUT THE PAPER: The development of big data in medical service networks provides research on the benefits of medical data, early diagnosis, patient care and network management. E-Health applications are especially important for patients who are not ready to see a specialist doctor or health care professional. The main purpose of the paper was to encourage physicians and families to predict disease using Machine Learning (ML) algorithms. In addition, many regions display different characteristics of certain important diseases, which can predict the prognosis of chronic diseases. The purpose of this study is to predict different types of chronic diseases using Gray Wolf optimization and auto encoder based Recurrent Neural Network (GWO + RNN). Usage was selected using GWO and chronic infections were predicted using the RNN method.

TRADE-OFFS OF THE PROPOSED SYSTEM

There are certain limitations in the proposed system which can be expected to be seen, they are:

1. While removing redundant values some important factors can be skipped which may give error in very certain cases.
2. Estimating some of the values can cause new biases to form in the system therefore to avoid such circumstances wither very rigorously complete dataset is required or such values with greater than 4% missing values have to omitted to prevent error.
3. Acquiring medical records is very difficult due to legal complications. Therefore, to avoid unnecessary legalities only small available test data can be used.

TECHNOLOGIES IMPLEMENTED

Prediction modelling is a process which uses datasets and probability for forecasting outcomes on the given input. A model is trained using a given dataset. We first process the dataset by replacing string values with numerical representation for easy manipulation, we then recognize empty values and replaces them with root mean square of the available values, then we remove the useless null values i.e. the values that are mostly zero. The dataset is then divided into two parts in a certain a:b ratio for training and testing purposes respectively. Once the data has been processed and divided the training set is then run through a classification algorithm which trains the model to predict the required output.

In this model we have used KNN as our classification algorithm as it provides the highest accuracy for the given set of data when compared to other classification algorithms. To apply KNN we first divide the train data into multiple parts then the initial data is represented on the graph and grouped on the basis of required result. We then follow three simple steps:

1. Calculating distance of given node from all other nodes.
2. Finding the K closest neighbors.
3. Voting for the maximum of which group are available and finding the mean and grouping with them.

It helps in grouping every element in its appropriate result. The experimental accuracy of the model varies with different disease areas depending on the complexity of the disease and its parameters involved. We have observed accuracy as high as 99% (for experimental data of Chronic Kidney Disease) and as low as 70% (for experimental data of Liver Disease).

CONCLUSION

The aim of this study is to help with early detection of Chronic Diseases which helps both the doctor and patient in management of the disease. The prediction plays an important role of identifying the indicators of a disease even when they are not prominent without the need for unnecessary invasive tests. This project works towards finding the optimal prediction model among a multitude of options available today and implementing the model with the highest accuracy with little to no bias in the system. It analyzes the limitations of the current model and builds on it to make the best system possible with highest real-world accuracy.

REFERENCES

[1] Shreekanth Jogar, Pavankumar Naik, Veeramma Vyapari, Madevi Vaddar, Kavita Dambal2, Bheemavva Hatti2"Chronic Diseases Prediction over Bigdata by using Machine Learning (Published in ." International Journal of Scientific Research in Computer Science, Engineering and Information Technology (2019). (BASE PAPER)

[2] Predicting individual disease risk based on medical history (Published in Proceeding CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management)

[3] An Analysis on Chronic Kidney Disease Prediction System: Cleaning, Preprocessing, and Effective Classification of Data by Ankit, Bhagyashree Besra and Banshidhar Majhi (Published in Recent Findings in Intelligent Computing Techniques, Advances in Intelligent Systems and Computing 707, https://doi.org/10.1007/978-981-10-8639-7_49)

[4] Smart collaboration framework for managing chronic disease using recommender system (Hussein, A., Omar, W., Li, X. et al. Health Syst (2014) https://doi.org/10.1057/hs.2013.8)Stock Market Prediction Using Machine Learning; 2018 First International Conference on Secure Cyber Computing and Communication(ICSCCC).

[5] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

[6] P. B. Jensen, L. J. Jensen, and S. Brunak.

[7] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H.

[8] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 855–864.

[9] S.-M. Chu, W.-T. Shih, Y.-H. Yang, P.-C. Chen, and Y.-H. Chu, "Use of traditional chinese medicine in patients with hyperlipidemia: A population-based study in taiwan," Journal of ethnopharmacology, vol. 168, pp. 129–135, 2015.

[10] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093, 2015