Predicting Diseases using Machine Learning: Fast Support Vector Machine and Stochastic Optimization Approach

Rajeev D M[1], Shadab Arshad[2]

[1]Student, Department of ISE, Sri Krishna Institute of Technology, Bengaluru.
[2]Student, Department of CSE, Sri Krishna Institute of Technology, Bengaluru.

Abstract - *Ailments that are related with the manner in which an individual or gathering of individuals live are known as way of lifestyle diseases. Healthcare industry gathers gigantic ailment related information that is shockingly not mined to find shrouded data that could be utilized for viable disease prediction. In order to make informed decisions for medical diagnosis and treatment choices, medical facilities need to be built. Machine learning in healthcare lets people analyze and interpret large and complicated medical datasets through clinical perspectives. Doctors could then use it further in the provision of healthcare. Machine learning, therefore, can result in increased quality of care when applied in healthcare. In the following paper we will be using a Fast support vector machine-based model for early disease prediction. Along with this we will be focusing on stochastic optimization of large medical datasets which will further help is in reducing the complexity as well as the costs.*

Key Words: Fast SVM, PCA, PLS, Stochastic optimization, Ailments.

1. INTRODUCTION

Predicting disease by implementing data analysis and machine learning methods using patient treatment record and health information is a constant struggle over the recent decades. For the analysis of various diseases, most researchers have implemented data mining strategies to clinical data or patient profiles. These methods have attempted to predict infection recurrence. In fact, some methods attempt to predict disease prevention and evolution. Way of life and diet are the two primary factors that are considered to impact Receptiveness to different maladies. Sicknesses are for the most part brought about by a blend of change, way of life choices, and environment. Likewise, distinguishing wellbeing dangers in a person's family is one of the most significant things an individual can do to support his/her professional comprehend and analyze inherited connected disorders like malignant growth, diabetes, and dysfunctional behavior. The primary spotlight is on to utilize AI in human services to enhance quiet think about better outcomes. AI has made simpler to distinguish various sicknesses and determination accurately. Prescient examination with the assistance of effective different AI calculations predicts the infection all the more accurately and help treat patients. Infections are mostly brought about by a mix of change, way of life choices, and environment. In addition, identifying wellbeing dangers in a person's family is one of the most pivotal things an individual can do to support his/her expert comprehend and analyze genetically connected disorders like malignant growth, diabetes, and dysfunctional behavior. Ailments that are related with the manner in which an individual or gathering of individuals live are known as way of life infection. For solving this problem, we would be using modified support vector machine. One of the major reasons behind using modified SVM is to improve its efficiency over original support vector machine model. We can improve our system by eliminating redundant vector spaces.

2. RELATED WORK

Lot of works has been done in the field of disease prediction by various scholars since last decade.

1) In Jingshu Liu, Zachariah Zhang and Narges Razavian's paper titled "Deep EHR: Chronic Disease Prediction Using Medical Notes" In this work the authors have proposed a general performance based various tasks system for sickness beginning forecast that consolidates both free-content restorative notes and organized data. They thought about execution of various profound learning designs including CNN, LSTM and progressive models. In difference to customary content based forecast models, their methodology doesn't require sickness explicit component building, and can deal with nullifications and numerical qualities that exist in the content. Their outcomes on an associate of around 1 million patients show that models utilizing content beat models utilizing simply organized information, and that models fit for utilizing numerical qualities and invalidations in the content, notwithstanding the crude content, further improves the execution. Also, they thought about various representation strategies for medicinal experts to translate model expectations. [1]
2) Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi and Leila Shahmoradi's paper "An analytical method for diseases prediction using machine learning techniques" In this paper, the approach taken is for information based framework for infections expectation utilizing bunching, clamor evacuation, and forecast procedures. They have used Classification and Regression Trees (CART) to create the fluffy principles to

be utilized in the information based framework. The testing of the proposed technique has been done on a few open medicinal datasets. Results on Pima Indian Diabetes, Mesothelioma, WDBC, StatLog, Cleveland and Parkinson's telemonitoring datasets show that proposed strategy strikingly improves the illnesses forecast exactness. The outcomes demonstrated that the mix of fluffy principle based, CART with clamor expulsion and grouping procedures can be successful in ailments expectation from certifiable medicinal datasets. The information based framework can help medicinal professionals in the social insurance practice as a clinical expository strategy. [2]

3) In Min Chen and Yixue Hao's paper titled "Disease Prediction by Machine Learning over BigData from Healthcare Communities" In line AI calculations for viable expectation of constant illness flare-up in infection visited networks. The authors have analyzed the altered expectation models over genuine medical clinic information gather from focal China in 2013-2015. To conquer the trouble of inadequate information, they have utilized an idle factor model to recreate the missing information, by probing a provincial ceaseless infection of cerebral dead tissue. A convolutional neural system based multimodal infection hazard expectation (CNN-MDRP) algorithm utilizing organized and unstructured information from hospital was proposed. To the best of their insight, none of the current work concentrated on the two information types in the territory of therapeutic huge information analytics. Compared to a few runs of the mill forecast calculations, the expectation precision of the proposed calculation arrives at 94.8% with an intermingling speed which was quicker than that of the CNN-based unimodal illness chance forecast (CNN-UDRP) calculation. [3]

3. BACKGROUND

1) SVM: Support Vector Machine (SVM) is a managed AI calculation technique which can be utilized for both order and relapse difficulties. It is a very important topic in the field of patter recognition, machine learning, bio informatics etc. SVM follows the approach of finding a hyperplane which maximizes the geometric margin and minimizes the classification error which is based on given a two class linear linearly separable variable [3]. In any case, it is for the most part utilized in characterization issues. In this calculation, plot every datum thing as a point in n-dimensional space where n is number of highlights you have with the estimation of each component being the estimation of a specific organize. Bolster Vectors are just the co-ordinates of individual perception. Support Vector Machine is an outskirt which best isolates the two classes. The support vector machine has been chosen because it represents a framework both interesting from a machine learning perspective and from an embedded systems perspective. An SVM is a linear or non-linear classifier, which is a mathematical function that can distinguish two different kinds of objects.

Training a SVM can be illustrated with the following pseudo code:

```
Algorithm 1 Training an SVM
Require: X and y loaded with training labeled data, a = 0
or a partially trained SVM
some value (10 for example)
repeat
    for all {$x_i, y_i$},{$x_j$,$y_j$} do
        Optimize $\alpha_i$ and $\alpha_j$
    end for
until no changes in $\alpha$ or other resource constraint
criteria met
Ensure: Retain only the support vectors ($\alpha_i > 0$)
```

2) Fast support vector machine: This approach is based on the theory of SVM, the properties of support vectors and a probabilistic formulation. As is well known, the objective of SVM is to find a hyperplane which provides the maximum margin for both classes. The margin between both classes is measured by the distances between support vectors of each class and the hyperplane. If a support vector s is not located close to the boundary of its associated class, we can find another training vector s' along the direction toward the hyperplane, and this vector s' is closer to the hyperplane than the support vector *s*, which means that *s* is not the support vector. As a result, one of the interesting properties of support vectors is that they are located at the boundary of their classes. According to this property, we propose to adopt the pruning strategy based on the Gaussian model to eliminate the training vectors which are close to the center of their associated classes. In addition, another interesting property of support vector is that they are located close to the hyperplane. As a result, the training vectors which are far away from the hyperplane can be eliminated by the pruning strategy based on the projection process. [4]

```
Algorithm FSVM (a set of training vector)
1. Eliminate training vectors by the Gaussian models
2. Eliminate training vectors by the projection process:
3. Perform SMO to obtain the binary SVM classifier
```

Proposed System:

We have divided our system in two major sub systems. Our first task would be to focus on collecting our data. We would like to create a real time system to collect data directly with our stakeholders which would be major hospitals, pathologies and other health institutes. The data preprocessing and feeding should be done in an automated way. Though real time system implementation are considered expensive, this would enable our system in order to achieve more accuracy during testing and training phase. The evolution of our system based upon this approach would be quite improving.

Once we are able to preprocess our raw data and obtain our dataset, we are focusing on reducing the dimensions of our dataset. This is majorly because of our size of the dataset. Reducing the dimensions has been a popular step in preserving the important features resulting in cost reduction. Various techniques being used are Principal Component Analysis, Partial Least square. We are implementing already existing optimizing techniques for this purpose which are stochastic in nature [5].

The second part of our design can be considered the major part of our system. We would be training our system based upon modified support vector machine also mentioned as Fast support vector machine [4]. Though the paper [4] has been implemented on image segmentation, we would be implementing it in terms of our medical dataset.

Based on a set of training vectors with suitable labels, the goal of the Fast Support Vector Machine Method (FSVM) is to remove spurious training vectors and to train the classifier through the residual training vectors. In certain terms, the distinction between FSVM and the current approaches to SVM is that FSVM focuses on reducing the number of vectors for redundant learning. There are two hypotheses which the authors [4] have considered important for the formulation of FSVM: there is a convex hull in each category for the input training vectors and the question must be separable. Next, by making the use of gaussian models, FSVM removes the training vectors that are close to the center of the class. Instead, through a projection method, it eliminates other training vectors. Eventually, in order to obtain the classification, FSVM performs sequential minimal optimization (SMO) on the remaining training vectors.

Methodology:

There would be four major steps in the methodology which are to be followed

1) Data Discovery: There are a lot of medical repositories which are present over the internet. Along with these one of our focus is towards collaborating with various hospitals and pathology labs. The enormous amount of raw data being processed daily should be processed precisely. Information disclosure can be seen as two stages. To start with, the created information must be filed and distributed for sharing. Numerous community frameworks are intended to make this procedure simple. In any case, different frameworks are worked without the intention of sharing datasets. For these frameworks, a posthoc approach must be utilized where metadata is produced after the datasets are made, without the assistance of the dataset proprietors. Next, someone else can look the datasets for their AI undertakings. Here the key difficulties incorporate how proportional the looking and how to tell whether a dataset is appropriate for a given AI task.

2) Data preprocessing: Here as our major source of data will be from Hospital and pathology labs, our focus Is on using real time data preprocessing systems. This technique can respond almost immediately to various signals to acquire and process information. These involve high maintenance and upfront cost attributed to very advanced technology and computing power. Once we have collected real time data, we can polish our data by removing noises, optimizing it. Other reason of this here would be to stochastically optimize our dataset which would be based upon non-convex optimization. As the dataset would be consisting of a lot of attributes, we can reduce the dimensionality of it. We can make use of Principal Component Analysis for this issue. Going one step ahead, we can also stochastically optimize it for reducing the complexity [5].

3) Training and testing: Under various circumstances, the proposed method needs training and tested by modifying FSVM variables in order to obtain correctness. Therefore, we find the reliability of the model to be optimum. From the data gathered, 80:20 would be used simultaneously to train and check the design. In event of necessity, arrangements on the algorithm used must be made to improvise. Apart from this our objective would be to reduce our dataset dimensionally so that we could extract and use the major features for our training and testing.

4) Result Analysis: The major factors being considered for the system is precision as well cost and complexity. Also, it would be determined how the system handles new data and processes it from the source. Apart from this the stochastic nature of modelling algorithm may determine how deterministic and accurate the results are.

CONCLUSION AND FUTURE WORK:

Disease prediction has majorly been considered as a magnificent field in Medical informatics. The earlier the diseases are predicted, magnificent are the chances of it being treated precisely. We feel that our modeled approach would be much more significant in working on real time data. Collecting this data would be a tedious task but would help us in making the system much more precise and helpful for the medical industry. Apart from this, there are systems which are based upon support vector machines for disease classification. But our approach for was to work on the heart of these systems aka Support Vector machines. Hence, Fast support vector machine suitably fitted into our model. Our future work would establish empirical results as well to focus on more practical deployment.

REFERENCES

[1] Liu, Jingshu & Zhang, Zachariah & Razavian, Narges. (2018). Deep EHR: Chronic Disease Prediction Using Medical Notes.

[2] Nilashi, Mehrbakhsh & Ibrahim, Othman & Ahmadi, Hossein & Shahmoradi, Leila. (2017). An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering. 106. 10.1016/j.compchemeng.2017.06.011.

[3] Chen, Min & Hao, Yixue & Hwang, Kai & Wang, Lu & Wang, Lin. (2017). Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2694446.

[4] Ivor W. Tsang , HK James T. Kwok , HK Pak-Ming Cheung Fast SVM Training on Very Large Data Sets. Published 4/05.

[5] Gupta, Mayank Raj. (2019). Understanding Stochastic Optimization of PCA, PLS and CCA Problem with a Focus on their Performance in Noisy Settings. 10.13140/RG.2.2.26365.41441.