Application of Machine Learning model in Diagnosis

Midway report

Group: Black Panthers

Webpage: https://williaddmw.github.io/DataMIningFinalProject.github.io/

## I. Introduction

Complex medical diagnosis is frequently a lengthy and complicated process, occupying prolonged amounts of time for clients, physicians and healthcare staff. Subsequently, medical resources are often inefficiently allocated, and client costs may be excessive. Increasingly, healthcare organizations are finding solutions to lengthy referral and diagnosis issues with machine learning and electronic surveys that serve as a medical triage to classify symptoms and refer clients to appropriate diagnosis recourse. Faster accurate diagnosis results in improved client outcomes, as well as saving time and money for all involved, and allows for more clients to be seen and helped by healthcare professionals.

BPPV (Benign Paroxysmal Positional Vertigo) is a common dizziness disease with very high misdiagnosis rate. Consequently, clients and healthcare professionals alike stand to benefit from methods for faster more efficient diagnosis of BPPV. This study aims to improve BPPV diagnosis by implementing machine learning methods to procure an electronic survey based clinical support system consisting of a minimum set of questions that can achieve an accuracy of approximately 90% for BPPV diagnosis.

Relevant studies have implemented a linear pre-visit questionnaire to expedite BPPV diagnosis (Friedland et. al), and more recently a study by Richburg, Povinelli and Friedland developed an electronic survey questionnaire as a clinical support system for BPPV diagnosis with high accuracy and specificity using machine learning. [1] In addition, medical implementation of machine learning models to help diagnose illness is seen in studies implementing a self- referral decision support framework for low back pain as in “Evaluation of three machine learning models for self-referral decision support on low back pain in primary care”, and work by Ayeldeen et. al., “Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach”. [2][3]

This project work is built on the framework of early two-phase survey conducted to diagnose BBPV using decision tree [2]. Two iterations of survey and data analysis have been conducted and our project is moving to the phase three and four of our research. Our study will employ decision tree and ANN models to generate and classify diagnostic features of BPPV and predict outcomes using Weka based on survey results. Machine learning models such as random forest, Naive Bayes, K-

Nearest Neighbors, Support Vector Machines in Python will compare decision trees and ANN models. Current accuracy rate using this model in diagnosis is around 70%. This study will implement the new models and examine accuracy for improvement, with considerations for the potential for certain noise in the data set. This project will continue to explore a novel ensemble approach to improve the accuracy of diagnosis for BPPV. Expectations are to develop a model with high accuracy, sensitivity and specificity in our prediction with the help of machine learning algorithms. Additional literary analysis will examine similar research to further understanding of potential methods to incorporate and deepen project understanding.

## II. Related Work

Relevant studies include work by Friedland et. al., which implemented a linear pre-visit questionnaire to expedite BPPV diagnosis. The questions were administered prior to the office visit and concerned patient history in an effort to garner enough information to allow for a narrowed differential diagnosis before the first clinic visit. Because the questionnaire was overly lengthy-10 pages- it was prone to questionnaire discrepancies i.e., skipped questions, dishonest reporting, lack of question meaning, etc. [4]

More recently a study by Richburg, Povinelli and Friedland, developed an electronic survey questionnaire as a clinical support system for BPPV diagnosis with high accuracy and specificity using machine learning. [3] The electronic survey allowed for keeping relevant questions, and the ability to pass over questions that were not relevant. Study objective was to employ machine learning models to classify features as BPPV or not and thereby providing clinical support and accelerating diagnosis and treatment.[3]

In addition, medical implementation of machine learning models to help diagnose illness is seen in studies implementing a self- referral decision support framework for low back pain as in "Evaluation of three machine learning models for self-referral decision support on low back pain in primary care", where, in an attempt to accurately prescribe the right intervention at the appropriate stage to circumvent condition attaining chronic status, specific machine learning models consisting of decision tree, random forest, and boosted tree techniques to classify low back pain cases.[2] Here, the boosted tree model performed best on the classification of low back pain cases, however, the evaluation measures confirm that all models provided referral advice better than just a random guess, meaning that all models learned some implicit knowledge of the provided referral advices in the training dataset. Finally, research by Ayeldeen et. al., "Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach", also used machine learning techniques to predict an individuals' degree of liver fibrosis. Here, using decision tree classifier techniques, researchers were able to achieve accuracy rates of 93.7%, higher than studies with similar conditions were able to attain. [2][1]

## III. Datasets and Features

### A. Description of dataset

The patient data is this study was collected from a survey questionnaire in Medical College of Wisconsin. The study coordinator in Medical College of Wisconsin would help the patients set up the tablets and assist them in filling out the survey in Android Tablets. The answers to the survey were stored in SQLite database when the survey was completed. The doctors in the hospital later would diagnose the patients if he or she has BPPV. Until the end of March in 2019, our research team has collected 74 patient surveys in phase one, 91 patient surveys in phase two, and 100 patient surveys in phase three. In total, there are 266 patient survey collected from three phases. The first phase of the survey contains 84 questions and began in May 2017. The second phase of the survey contains 68 questions and began in November 2017. [1] The third phase of the survey contains 42 questions and begin in August 2019 and is still going on. The survey contains six sections of questions with focus on symptoms of dizziness, timing of attacks, feeling of ear pain and headaches as well as triggers for dizziness and medical history. The phase two survey contains supplemental section. The supplemental section contains six questions and is used to test how well two decisions trees which are generated based on phase one survey would perform.

### B. Combination of three survey questions

The project has implemented three versions of survey in three phases. Each version of survey has same overlapped questions with the other two. There are 84, 74, and 42 survey questions respectively in phase one, two, and three surveys. In order to efficiently use all of the patient data, we match the same questions from all three phases and get a combined data set.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | P1: questionnumber | P1:questiontext | Matching P2: questio | Matching P3 questionnumber | | |
| 2 | symptoms_q1a | 1a) Although you may experience many sensations, what is the single most noticeable part of y | natureq1(1) | natureq1(1) | | |
| 3 | symptoms_q2 | 2) Has your dizziness occurred once or more than once? | natureq6(6) | natureq3(3) | | |
| 4 | symptoms_q5 | With your responseTo1a have you had nausea and/or vomiting? | natureq8or9(8,9) | nature5(5) | | |
| 5 | symptoms_q6 | With your responseTo1a have you had double vision? | natureq11(11) | nature7(7) | | |
| 6 | symptoms_q7 | With your responseTo1a have you had blurry vision? | natureq10(10) | nature6(6) | | |
| 7 | duration_q1 | 3) Is your responseTo1a currently with you 24 hours a day, never stopping? | nature5(5) | nature2(2) | | |
| 8 | duration_q3a | 5a) Does your responseTo1a last seconds to 1 minute? | temporalq2(15) | temporal1(8) | | |
| 9 | duration_q3d | 5d) Does your responseTo1a last about one hour? | temporalq4(17) | temporal2(9) | | |
| 10 | duration_q3e | 5e) Does your responseTo1a last hours but less than 12 hours? | temporalq5(18) | temporal3(10) | | |
| 11 | duration_q3g | 5g) Does your responseTo1a last 1 day or longer? | temporalq7(20) | temporal4(11) | | |
| 12 | yesno1_q4 | Is your responseTo1a typically made worse or triggered by lying down or rolling in bed? | trigger7or8(27,28) | trigger4or5(15,16) | | |
| 13 | yesno1_q7 | Is your responseTo1a typically made worse or triggered by automobile rides? | trigger15(35) | trigger10(21) | | |
| 14 | yesno1_q8 | Is your responseTo1a typically made worse or triggered by loud sounds? | trigger14(34) | trigger9(20) | | |
| 15 | yesno1_q12 | Is your responseTo1a typically made worse or triggered by sitting up or standing up? | trigger4(24) | trigger1(12) | | |
| 16 | yesno1_q13 | Is your responseTo1a typically made worse or triggered by walking on uneven ground? | trigger16(36) | trigger11(22) | | |
| 17 | yesno1_q15 | Is your responseTo1a typically made worse or triggered by supermarket aisles, malls, or tunnels? | trigger 12or13(32,33) | trigger7or8(18,19) | | |
| 18 | yesno1_q17 | Is your responseTo1a typically made worse or triggered by turning your head while walking? | trigger11(31) | trigger6(17) | | |
| 19 | yesno1_q18 | Is your responseTo1a typically made worse or triggered by driving a car at night? | trigger17(37) | trigger12(23) | | |
| 20 | yesno1_q19 | Is your responseTo1a typically made worse or triggered by reaching or bending? | trigger5(25) | trigger2(13) | | |
| 21 | yesno2_q4 | Have you had a total of 5 or more bad headaches in your lifetime? | migraine1(56) | migraine2(31) | | |
| 22 | yesno2_q11 | Have you ever had a headache that throbs or pulses? | migraine3(58) | migraine3(32) | | |
| 23 | yesno2_q13 | Have you ever had nausea or vomiting with a headache? | migraine7(62) | migraine8(37) | | |
| 24 | yesno2_q14 | Have you ever had increased sensitivity to light with a headache? | migraine5(60) | migriaine5(34) | | |
| 25 | yesno2_q15 | Have you ever had increased sensitivity to sounds with a headache? | migraine6(61) | migraine6(35) | | |
| 26 | yesno2_q17 | Have you ever had your responseTo1a associated with a headache? | migraine8(63) | migraine10(39) | | |
| 27 | ear_q5 | Do you have ringing or other noise in your ears (tinnitus)? | ear15or16(52,53) | ear4or5(27,28) | | |
| 28 | ear_q7 | Do you have pain in your ears? | ear11or12(48,49) | ear2or3(25,26) | | |
| 29 | ear_q8 | Do you get frequent ear infections? | ear10(47) | ear1(24) | | |
| 30 | yesno3_q12 | Have you had a hip or knee replacement? | history4(67) | history2(42) | | |
| 31 | | | | | | |
| 32 | | | | | | |

## IV Methods

Classification using machine learning methods. The following machine learning models will be implemented in this study:

1. Decision trees

The implementation of decision trees model consists of two steps. The first step is to obtain a subset of the most relevant questions using the correlation attribute evaluator in the Weka. The threshold of correlation attitude evaluator is 0.25. Then a J-48 decision tree algorithm in Weka will be used to classify each record as "Has BBPV" or " No BPPV".

2. Ensemble of decision trees

The project also explores a novel ensemble approach to improve the accuracy of diagnosis for BPPV. This method will generate multiple trees using training data and make predication based on majority vote and weighted decision. Therefore, a group of trees instead of individual tress will be combined to predict BPPV disease.

3. ANN model

ANN model is another machine learning framework that is inspired by biological neural network. The model takes features as input and results as output. There would be hidden layers called perceptron to process the information. We plan to use Weka to adjust the number of layers, the number of perceptron and learning rate to achieve the best prediction we could have.

4. KNN

KNN was a method of choice as the model structure is decided from the data- an outcome this study is interested in. This model makes predictions by calculating an input similarity to a training instance. [6] Not wanting to hold authentic medical data to assumptions, a goal of this research is to determine which survey questions provide the highest accuracy for BPPV diagnosis. KNN is often a go to classification method when prior knowledge about data distribution is scant. [6]

5. Naïve Bayes

Naive Bayes classification is based on Bayes theorem which provides us a way to calculate the probability of our classification based on existing data.

[1]Bayes' Theorem is stated as:

$P(h|d) = (P(d|h) * P(h)) / P(d)$

Where

- **P(h|d)** is the probability of hypothesis h given the data d. This is called the posterior probability.

- **P(d|h)** is the probability of data d given that the hypothesis h was true.
- **P(h)** is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- **P(d)** is the probability of the data (regardless of the hypothesis).

You will notice that we are trying to calculate the conditional probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h). Once we calculate the probability of a different number of hypothesis then we select the hypothesis with the highest probability.

Naive Bayes theorem is called Idiot Bayes because of the probability calculation of each hypothesis is made simpler. It is assumed in this theorem that features are independent of each other and do not have relation to each other what so ever.

This is a very strong assumption and might not be true with most of the real-time data. However, this theorem seems to work well with most of the cases. [5]


[1]H. Ayeldeen, O. Shaker , G. Ayeldeen, and K. M. Anwar, "Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach∗," *IEEE*, Nov. 2015.

[2]W. O. Nijeweme-d'Hollosya, L. van Velsen, M. Poel, C. G. M. Groothuis-Oudshoornd, R. Soer, and H. Hermens, "Evaluation of three machine learning models for self-referral decision support on low back pain in primary care," *International Journal of Medical Informatics*, vol. 110, no. 31, Feb. 2018.

[3]H. A. Richburg, R. J. Povinelli , and D. R. Friedland , "Direct-to-Patient Survey for Diagnosis of Benign Paroxysmal Positional Vertigo ," *2018 17th IEEE International Conference on Machine Learning and Applications* , Dec. 2018.

[4]D. R. Friedland, S. Tarima, C. Erbe, and A. Miles, "Development of a Statistical Model for the Prediction of Common Vestibular Diagnoses," *JAMA Otolaryngology Head Neck Surgery*., Apr. 2016.

[5] J. Brownlee, "Naive Bayes for Machine Learning," *Machine Learning Mastery*, 11-Apr-2016. [Online]. Available: https://machinelearningmastery.com/naive-bayes-for-machine-learning/. [Accessed: 04-Apr-2019].

[6] A. Bronshtein, "KNN makes predictions just-in-time by calculating the similarity between an input sample and each training instance.," *Noteworthy- The Journal Blog*, 11-Apr-2017. .