Ali Rizvi, Alwin Baby, Kelly Holtzman, and Mansi Patel

Faculty of Engineering and Applied Science, University of Regina

ENSE 496AC: Artificial Intelligence

Dr. Kin-Choong Yow

October 30, 2020

Project Progress Report

Project Title. Medi-App: Self-Diagnosis of Illnesses

Project Summary

Medi-App attempts to provide the symptom-to-illness investigation potential of online clinical databases without the time consumption of searching online. Symptoms and illness prognosis data are from online machine learning databases under license allowing academic use. The data are then used to initially train a Bayesian Network alongside an additional supporting workflow as suggested by related studies. The resulting probabilistic model(s) provide Medi-App with the capability to compare the user's symptoms with past observed symptoms, and report the probability that the user is observing symptoms that indicate the presence of a particular illness or disease.

Problem Description

The outbreak of the disease COVID-19 has drawn the common person to reflect on symptoms of illness more frequently. Symptoms similar to the coronavirus are being investigated more on search engines as a result (Google Trends, n.d.). The medical world is vast, and the average person may find themselves down a rabbit hole of symptoms and illnesses that may not even relate to what they were investigating in the first place. To prevent the potential situation where the user ends up convincing themselves they have an unlikely illness, our team proposed to create an application for self-diagnosis of illnesses that would provide legitimate suggestion(s) for the mystery illness without the user needing to research on their own.

Data and Algorithms Anticipated

Legitimate clinical data in the format of real samples of symptoms and prognosis are hard to find without connections to medical professionals who can provide such a dataset without personally-identifying information; as such, our team has investigated online

machine learning databases including the UC Irvine Machine Learning Repository and Kaggle which provide cleaned datasets for academic purposes. We have tentatively chosen a dataset for disease prediction with samples of 132 possible symptoms and related prognosis of 42 illnesses from an anonymous student representing Nirma University (2020), such a dataset represents the format of symptoms-to-illness we would require for our project. We recognize the validity of such a dataset is questionable and are investigating further at this time.

Our team has investigated a variety of artificial intelligence algorithms that would help us to both model the distribution of symptoms to illnesses and predict the most likely illness given the set of user symptoms: an Artificial Neural Network (ANN) using the `scikit-neuralnetwork` library on top of the `scikit-learn` python module, k-means clustering using `scikit-learn`, and a Bayesian Network (Bayes net) using the `pomegranate` python library. Each of the algorithms listed utilize statistical probability relationships to reduce prognosis suggestions according to our chosen dataset for training, which is exactly what we intend Medi-App to do for the user.

Initial Results of Investigation

After logically analyzing the three algorithms we found that the first best one to model symptoms to illnesses was the Bayes net because of the model's inherent ability to infer the presence of a missing node in it's graphical representation. The article by Auras, et. al (2016) studies the use of Bayes nets in breast disease diagnosis and suggests, given legitimate training data for the model and the support of medical professionals in confirming the relationship between symptoms and disease, that Bayes nets could successfully be used to diagnose illnesses. We ought to be able to improve the performance of the Bayes net if we retrain the network on new input, and maintain the network's probability distributions from the results of first approximate inference.

The other models we investigated also had studies that suggested their use for disease diagnosis, should the model be combined with an algorithm with the capability to predict the disease (for our project, such a role could be fulfilled by Bayes net): particularly Vijayalakshmi, et. al (2020) recommend a workflow to be used with k-means clustering with a pre-processing phase much like the post-processing phase suggested by Auras, et. al (2016); and the conference proceeding by Neves, et. al (2015) suggests the data alone are not enough information, hence the need for medical professional opinion on the data used for training.

Project Future Direction

Our direction for the rest of the semester will be to implement our best model, the Bayes net, and integrate the recommendations set out by Auras, et. al (2016). There is also the possibility that we could extend the other two algorithms we investigated with the initial results of the Bayes net implementation, and form a better solution. We'll be critically evaluating the implementation's accuracy with test data, as suggested by our chosen dataset; and the run time of the implementation: we would like to present an application that does not force the user to wait longer than should be acceptable, within a time frame discussed by the team. Our results from implementation and reflection will suggest a model that should be used in future applications that aim to solve the problem of disease diagnosis with artificial intelligence. We have been documenting our project's progress and results in a GitHub software repository of the same name, see the Appendix for further details.

Individual Team Member Achievements

Each team member has completed the following responsibilities thus far (please refer to our project proposal for the full set of responsibilities set out to be completed for the Final Report):

1. (Kelly Holtzman) Completed investigation of the state-of-the-art of Bayes net and determined that the algorithm is fit for our purpose
2. (Mansi Patel) Completed investigation of the state-of-the-art of ANN and determined that the algorithm is fit for our purpose if used with another predictive algorithm
3. (Ali Rizvi) Completed investigation of the state-of-the-art of k-means clustering and determined that the algorithm is fit for our purpose if used with another predictive algorithm
4. (Mansi Patel) Completed initial investigation of the application component responsible for gathering legitimate symptom-to-illness data from reputable clinical databases; which became the dataset by Anonymous (2020).
5. (Alwin Baby) Completed implementation of the application component responsible for gathering and cleaning the user's input symptoms

Our GitHub repository, linked in Appendix A, explicitly details the documentation and implementation details by each team member.

References

Auras, H., Merouani, H. F., & Refai, A. (2016). Maintenance of a Bayesian network: application using medical diagnosis. *Evolving Systems*, 7(3), 187-196. doi: 10.1007/s12530-016-9146-8. https://casls-primo-prod.hosted.exlibrisgroup.com/permalink/f/1ed24l6/TN_cdi_crossref_primary_10_1007_s12530_016_9146_8

Anonymous (Nirma University) (2020). Disease Prediction Using Machine Learning (Version 1) [Data set]. Kaggle. https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning

Google Trends. (n.d.). *Coronavirus Search Trends.* Google. https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en

Neves J., Cunha, A., Almeida, A., Carvalho, A., Neves., J., Abelha, A., Machado, J., & Vicente, H. (Eds.) (2015). Artificial Neural Networks in Diagnosis of Liver Diseases. *Information Technology in Bio- and Medical-Informatics (ITBAM 2015)*. https://doi.org/10.1007/978-3-319-22741-2_7

Vijayalakshmi, S., Pallawi, Shruti, & Genish, T. (2020) Alzheimer Disease Detection Using Edge Enhanced K-Means Clustering Algorithm. *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*. http://dx.doi.org/10.2139/ssrn.3545092

Appendix

Our team has begun a GitHub software repository to track our progress and make our study publicly available. Please find the repository at the following:

Baby, A., Holtzman, K., Patel, M., and Rizvi, A. (2020). Medi-App. GitHub. https://github.com/holtzmak/Medi-App