Your Project Title Here

Juan Aceron, Marc Teves, Martin Thomas
Department of Computer Science
College of Engineering
University of the Philippines, Diliman

***Abstract*—Assessing the severity of heart disease in a patient is important. Machine learning algorithms can solve this problem by considering risk factors and their correlation with the presence of heart disease. Most machine learning algorithms suffer the curse of dimensionality, wherein the time taken to fit a model with training data suffers as the number of features increases. In addition, having more features tends to overfit the model with the data. To avoid the problems with excessive features, feature selection, also known as feature ranking, is employed to reduce features in datasets with large amounts of features. The group uses the naive approach of feature selection, evaluating the performance of each element in the powerset of the feature set. Performance is taken as the weighted average of each non-diagonal cell in a confusion matrix, so that more severse false negative cases have a higher weight. In the case of assessing heart disease, each feature is extracted using medical procedures with varying cost. This is taken into account when selecting the best feature set, so as to obtain the feature set that is most effective computationally, and in terms of cost. The group found that (feature set) is the most effective in terms of performance, while (feature set) is the most cost effective.**

I. Introduction

Heart disease, particularly coronary heart disease (CHD), is the leading cause of death among adults in the industrialized world [2]. CHD and stroke are life threatening conditions which require expensive treatments when they occur. When diagnosed correctly and prevented, the preventative measures for CHD and stroke do not cost much. If heart disease is prevented from progressing, there will be a lot of money saved in medical treatments.

Zethraeus et. al.[3] estimated that in 1999 in Sweden, preventing the onset of CHD or stroke saved each patient a medical bill ranging from 36 000 SEK to 91 000 SEK. In today's money, that would range from 274348 Php to 693498 Php per stroke prevented. [4]

Preventing and accurately diagnosing heart disease risk factors translates to major savings in medical fees. Getting checked for heart disease is the same as almost any other medical diagnosis. The patient is subjected to various non-invasive medical tests with varying prices. For example, finding the patient's cholesterol levels will cost 7.27 CAD, while finding the maximum heart rate achieved through a thalium stress test costs 102.90 CAD. [5]

After receiving the tests, a trained medical professional decides whether or not the patient has heart disease, and the severity of said heart disease.

Machine learning approaches have been proposed to automate the task of diagnosing heart disease.[5] [6] Ultimately, this is a data analysis problem. A common theme with data analysis is the *curse of dimensionality*. As the number of features grows in a dataset, the predictive power of the model used to characterize it increases up until a point where it starts to decrease. [7] With the case of machine learning, the time taken to fit the model with the data increases, for diminishing returns in performance. It would be helpful if only the most relevant and most predictive features are selected for data analysis.

This problem can be solved with *feature selection*, or *feature ranking*. Feature selection is the task of selecting the best subset of features from a given dataset that will create the best predictor, in terms of performance.

Medical diagnosis problems will benefit not only in training time and accuracy of their predictive models, but in saving diagnosis cost as well, since the number of tests will be reduced for each patient.

In this project, the best performing subset of features will be selected using naive feature selection. The subset of features will be evaluated by their performance in a support vector machine (SVM).

II. Short of Review of Related Studies

The Cleveland heart disease dataset is a well-known dataset used for the heart disease diagnosis problem. Many SVM models have been created to analyze this dataset.

Khanna et. al. [6] created an SVM model with an F-score of 0.87. However they did not provide a detailed methodology. They also did not do any feature selection.

Nahar et. al. [8] used two techniques of feature selection: manual (MFS) and computerized (CFS).

Without feature selection, their SMO classifier got an F-score of 0.862. With MFS, it got an F-score of 0.8. With CFS, it got an F-score of 0.815. Taking the intersection of feature subsets (MFS+CFS), the SMO classifier got an F-score of 0.861.

Compared to the full feature set, the SMO model seemed to do worse after MFS+CFS. The loss in performance is insignificant compared to the savings in medical fees by taking the reduced 3-feature subset as a basis for predictive models, instead of the 14-feature feature set. Nahar concluded that chest pain type, maximum heart rate, and exercise induced angina are the best features to use to decrease training time and decreasing medical cost. This study points to the idea that reducing the feature set will not incurr significant losses in performance.

III. Methodology and Results

The dataset used for the SVM classifier was the Cleveland heart disease dataset. The dataset has

1) wew

Discuss the dataset, preprocessing, machine learning technique you used, training parameters set, performance measure and validation method. Include tables, flowcharts, and figures, as necessary.

Discuss the performance of the model. Do the analysis here. Justify the setup. Show the experimental values vis-a-vis parameter values. Contrast with results of previous studies, if any.

Discuss some specific cases of pitfalls (misclassification, errors) and possible reasons behind them.

IV. Conclusion

The conclusion goes here. What were you able to accomplish? Were there any significant improvements from the previous studies reviewed? Were you able to build a model to address the topic?

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] American Heart Association. *Heart Disease and Stroke Statistics2003 Update*. Dallas, Tex: American Heart Association; 2002.

[3] Zethraeus N, Molin T, Henriksson P, Jnsson B. *Costs of coronary heart disease and stroke: the case of Sweden*. J Intern Med. 1999;246(2):151-9.

[4] Swedish Consumer Board. *CPI, Fixed Index Numbers (1980=100)*. Statistics Sweden. 2018

[5] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology*, *64*,304–310.

[6] D. Khanna, R. Sahu, V. Baths, and B. Deshpande. *Comparative Study of Classification T echniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease*. International Journal of Machine Learning and Computing, vol.5, no. 5, pp. 414-419, 2015.

[7] Trunk, G. V. (July 1979). *A Problem of Dimensionality: A Simple Example*. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (3): 306307.

[8] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y.-P. P. (2013). *Computational intelligence for heart disease diagnosis: A medical knowledge driven approach*. Expert Systems with Applications, 40(1), 96104. https://doi.org/10.1016/j.eswa.2012.07.032