Alex Teboul

Paper Review 2

DSC 540: Advanced Machine Learning
Professor: Casey Bennett

Paper: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes
Authors: Wei Yu*, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury

Review

Researchers Wei Yu et al. studied the application of Support Vector Machine (SVM) modeling towards diabetes and pre-diabetes classification. Their aim was to demonstrate that SVMs can be used effectively towards the classification of disease types, with diabetes as their case study. Specifically, they wanted to prove that SVMs could distinguish between individuals with diabetes or pre-diabetes and those without the disease at least as well as one of the most supported methods at the time - Logistic Regression. In order to accomplish this, they used data from the 1999-2004 National Health and Nutrition Examination Survey (NHANES) to classify diabetes using SVMs and Logistic Regression under two different Classification Schemes. They then compared the results between SVM and Logistic Regression models for the two Classification Schemes using the ROC Area Under the Curve (AUC) metric, sensitivity, specificity, PPV, and NPV. The researchers ultimately determined that SVMs are an effective modeling approach to classify diabetes. They also built a web-app that demonstrates their model by asking users to enter values for 8 or 10 of the relevant variables, and presenting a classification as diabetes or not diabetes. Methods like the one proposed here by Yu et al., do not require lab tests and can serve as useful tools to warn individuals about their diabetes risk.

Back in 2010 when the paper was released, an estimated 23.6 million Americans had diabetes, with an additional 57 million likely living with prediabetes [1]. Prediabetes is a term used to refer to individuals at high risk of developing diabetes as indicated by risk factors such as elevated blood glucose levels. Diabetes in general is a disease characterized by a failure of the body to regulate insulin production, leading to elevated blood sugar levels and a host of complications. Ultimately, diabetes is a serious disease, in terms of diminished quality of life, mortality, and expense. As of 2018, estimates place the number of diagnosed diabetes cases in the U.S. at 26.9 million, with 88 million living with pre-diabetes [2]. While the researchers were not trying to solve the problem of diabetes, this does demonstrate the relevance of study, as the problem of diabetes in the US has only gotten worse since this paper was released.

Data

The 1999-2004 National Health and Nutrition Examination Survey (NHANES) dataset is a cross-sectional, probability sample survey of the U.S. population. This representative study includes detailed demographic, health history, and behavioral information on participants. The researchers used non-pregnant participants over the age of 20 for their survey and made use of the following 14 feature variables: family history, age, gender, race and ethnicity, weight, height, waist circumference, BMI, hypertension, physical activity, smoking, alcohol use, education, and household. Determination of diabetes class (diabetes, pre-diabetes, diabetes but undiagnosed, and no diabetes) was based on survey participant responses. Specifically, if they had been diagnosed by a doctor or they had fasting plasma glucose in ranges associated with diabetes or prediabetes. The dataset included the following sample counts: Diagnosed diabetes (1,266), Undiagnosed diabetes (195), Pre-diabetes (1,576), No diabetes (2,277). A total of 6,314 samples were included for modelling, with 80% used for training and 20% for validation/testing. This dataset has some imbalances, but not so much so where it would very impact the findings of the paper.

Automatic feature selection was used to determine important features for the SVM and Logistic Regression models. In terms of the features that are relevant to this classification, they determined that 8 variables were sufficient predict if someone had diabetes (diagnosed or undiagnosed) versus those without diabetes (no diabetes or prediabetes) for Classification Scheme I. These variables were family history, age, race, weight, heigh, waist circumference, BMI, and hypertension. Under Classification Scheme II, which was to predict between those with diabetes who were undiagnosed or pre-diabetics and those who definitely did not have diabetes, they found 10 variables to lead to optimal performance of the classifier. These features were family history, age, race and ethnicity, weight, height, waist circumference, BMI, hypertension, sex, and physical activity. In general, these features make sense, as incidence and risk of diabetes are closely related to demographic and basic bodily health features. After normalizing the values of the different features, the researchers used LibSVM and SAS-callable SUDAAN v9 to train the SVM and Logistic Regression models.

Methods

The main machine learning method explored in this paper is Support Vector Machines. SVMs perform classification by constructing multidimensional hyperplanes to discriminate between two classes. They accomplish this by maximizing the margin between data clusters and make use of kernel functions to transform the input space. Linear, polynomial, RBF, and sigmoid kernels were tested in this paper, with Linear, RBF,

and Sigmoid performing comparably and polynomial performing worse.

The performance of the SVMs was compared to Logistic Regression models using the same features for Classification Scheme I and II. Logistic Regression, for reference, is a simple statistical method which uses a logistic function to model a binary variable. Both SVMs and Logistic Regression have been classically used for binary classification tasks.

In terms of general results of the paper, the AUC for the detection of diagnosed diabetes or undiagnosed diabetes was 83.47%, and it was 73.18% for pre-diabetes or undiagnosed diabetes in the validation test[1]. These best models were SVM with RBF kernel and SVM with linear kernel respectively. Though the SVMs with linear, RBF, and sigmoid performed almost identically. These results were compared AUCs for logistic regression and it was found that Classification Schemes I and II were 83.19% and 73.35% AUC respectively. They determined there was no statistically significant difference in the scores between SVM and Logistic Regression. Test, training, and 10-fold cross validation evaluation metrics for sensitivity, specificity, PPV, NPV, and AUC were also produced for the SVM models on the two Classification Schemes.

Critiques - The Good and the Bad

On the whole, I can find very few critiques of this paper. I like the way it was laid out and presented in terms of the methods used, clarity of methods explanation, and conclusions drawn by the researchers. The presentation of what the researchers planned to do was clear: compare SVM and Logistic Regression classification performance on a diabetes dataset using two different classification schemes. The results were clear as well and multiple metrics were displayed for AUC, sensitivity, specificity, etc of the SVMs. They demonstrated that there was no statistically significant difference in the AUCs obtained via SVM and Logistic Regression for the dataset. Their conclusion was well supported, that the SVM model performance was at least equivalent to the supported epidemiological method of Logistic Regression. They followed up their research by creating a useful web-tool that demonstrated the model in action. While this is certainly not equivalent to a diagnosis, it could be useful as an awareness measure, which is important given the prevalence of pre-diabetes and undiagnosed diabetes in the United States.

If I have one main critique, it is that the researchers did not elaborate on why the SVMs and Logistic Regression performed similarly. For example, they highlighted the SVM with RBF kernel as the best model for CS-I at 0.8347 AUC, but the SVM with linear kernel had an AUC of 0.8332. These do not appear to be significantly different, just as the SVM-RBF and Logistic Regression were not statistically significantly different. Often, Logistic Regression and Linear SVMs will perform similarly on low-dimensional space datasets like the one explored in this paper. It is likely that the decision boundaries

created by the SVMS and Logistic Regression are similar in this case. While SVMs have been shown to offer improved performance in a variety of fields, I would argue that for equivalent performance in this dataset, the simpler model of Logistic Regression might not need replacement by SVM for the purposes of diabetes classification by questionnaire. That said, the researchers are not claiming that SVMs should replace Logistic Regression. They simply highlighted SVMs as a relevant modeling technique for diabetes classification and called for further tests/studies of its applicability towards other common disease classification tasks.

Conclusion

The researchers tested two classification schemes to detect diabetes and pre-diabetes on a dataset for the U.S. population. They found that Support Vector Machines have relatively good performance on the dataset, highlighting the technique's potential for use in classifying other common diseases. It was also determined that SVMs had equivalent performance to the more commonly accepted method of Logistic Regression, at the time when the paper was released. The researchers took their model a step further by releasing a web-app that allows users to respond to questions and get classified using the SVM model presented in the paper for diabetes. The web-app questions are based on the same features that were relevant to the SVM model in the paper, namely: family history, age, race and ethnicity, weight, height, waist circumference, BMI, hypertension, sex, and physical activity. None of these features require lab-testing, and most individuals can determine answers to the questions, making it a useful tool.

I believe the researchers did a good job of explaining their methods and had well thought out conclusions. They also called for further study and did not overblow expectations for the use of SVMs. Their approach also supports potential for awareness measures that allow individuals to get warned of their diabetes risk via questionnaire. Early awareness with chronic diseases like diabetes that have multiple lifestyle risk factors is very important. Since its release in 2010, this paper has been cited 227 times, highlighting its influence in the space of diabetes and machine learning research.

Citation:

[1] Yu et al.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making* 2010 10:16.

[2] Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020.