Characterizing Atrial Fibrillation Burden for Stroke Prevention

Lichy Han

Atrial fibrillation (AF) is an arrhythmia that is common, affecting nearly 3 million in the US. Although AF is known to cause stroke, conventional biostatistical approaches have failed to robustly predict stroke from continuous AF burden data. I explored stroke prediction from AF patterns by using supervised machine learning in a large cohort of cardiac implantable electronic device patients remotely monitored in the Veterans Administration Health Care System. AF burden in the 30 days prior to ischemic stroke in cases were compared to the first 30 days of monitoring in matched non-stroke controls using a variety of processed feature sets. The best performing set and test combination was then used to identify the optimal number of days to analyze. Using dynamic time warping with K-nearest neighbors on a time interval of 60 days was found to be the highest performing based on a F1 measure of nearly 0.7, specificity of 0.51, and training and test errors of roughly 0.35, demonstrating that this still remains a challenging problem in time series classification.

Introduction

Atrial fibrillation (AF) and atrial flutter are arrhythmias of the atria that are known to be a cause of ischemic stroke when present in the left atrium. AF is exceedingly common, and resulting strokes can be prevented with proper management and anticoagulation. However, though atrial arrhythmias are common, not all of them lead to thrombotic events such as stroke, and it is not known yet what the treatment threshold for atrial fibrillation should be. AF is thought to increase the risk of stroke by 5-fold, and strokes as a result of AF are thought to be more harmful than strokes not associated with AF.[1] At the same time, it is not known if there are changes that happen to the pattern and amount of AF prior to stroke. Identifying such patterns would allow for improved, real-time prediction of stroke and affect anticoagulation strategies for AF patients. Classically, it is thought that any amount of AF will increase the risk of stroke, and previous studies such as the ASSERT trial showed a correlation between subclinical AF (AF lasting at least 6 minutes) and stroke.[2] In this study, I further explore the temporal relationship between AF and stroke, using time series analytical techniques and supervised machine learning.

Methods

Study Subjects

A total of 9836 patients from the Veterans Administration Health Care System were enrolled in this study, and all patients enrolled gave written, informed consent prior to entering the study. Patients had either an implantable cardioverter defibrillator or pacemaker with an atrial lead. Remote monitoring using the atrial lead from these devices reported the number of seconds of AF a patient had per day, also known as the AF burden. The mean follow-up was 934 days, with the maximum at 2765 days. Remote monitoring was terminated in the event of patient death, patient censoring, or stroke.

Data Preprocessing

Patients who had an ischemic stroke were matched to a control patient based on several clinical variables including gender; history of myocardial infarction, heart failure, hypertension, diabetes mellitus, and other co-morbidities; scores on the Selim comorbidity index and the Charlson comorbidity index; and the $CHADS_2$ clinical risk score for stroke. For cases that had multiple perfectly matched controls, the control with the closest age was selected. Intervals of length *L* = 15, 30, 60, 90, and 120 days were selected as the last *L* days before stroke in cases and the first *L* days of remote monitoring in controls, respectively.

Data were processed, yielding 4 sets of features: 1) raw data without any processing, 2) data normalized to have zero mean over the interval length for each patient, 3) difference data, where each data point represents the AF burden on day *i+1* minus the AF burden on day *i* for *i* from *1* to *L-1,* and 4) data processed using the fast Fourier transform (FFT).

Supervised Machine Learning

All analyses were done using R 3.0.2. Due to the small number of subjects in the case control subgroup, all classification methods were applied using 5-fold cross validation.

K-Nearest Neighbors (KNN)

KNN is a non-parametric method that classifies each sample based on its k-nearest neighbors. KNN was implemented using the class package using Euclidean distance with k=3 and also implemented using dynamic time warping with the dtw library and k = 1.[3,4] Dynamic time warping is a dynamic programming method used in time series analysis to align two time series and calculate the similarity between them. This distance metric was then used in KNN to determine the nearest neighbor(s).

Logistic Regression

Logistic regression was performed using the LiblineaR package, with L2 regularization and a penalty of C = 1.0.[5]

Support Vector Machines (SVMs)

An SVM model was created using the e1071 package. The model uses a Gaussian kernel, with a penalty of C = 1.0 and gamma as 1/(number of features). [6]

Random Forest

Random forest is an ensemble method using multiple decision trees. Using the randomForest package, I applied a random forest classifier using 2 trees.[7]

Results

Of the 9836 patients, 306 had a stroke during the remote monitoring period. At interval lengths of 15, 30, 60, 90, and 120 days, there were 56, 64, 76, 44, and 48 subjects, including cases and controls, in each analysis respectively. Cases that were not included either did not have the required length of consecutive monitoring prior to stroke, had no atrial fibrillation in the specified interval given prior to stroke, or did not have a suitable matched control. An example of the AF burden in cases and controls on the first day of the selected 30 day interval is shown in Figure 1.

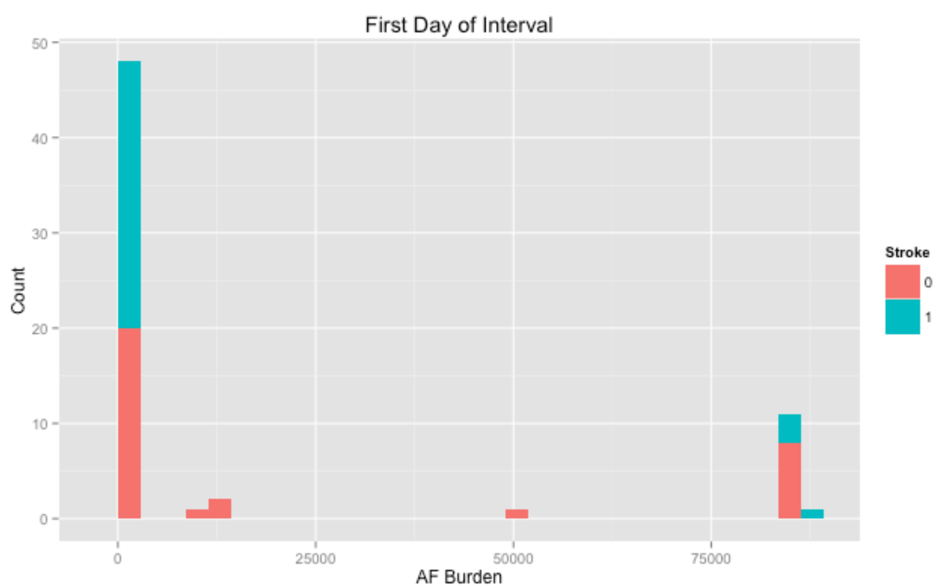

Figure 1. AF burden for the first day of monitoring in cases (Stroke = 1) and in controls (Stroke = 0)

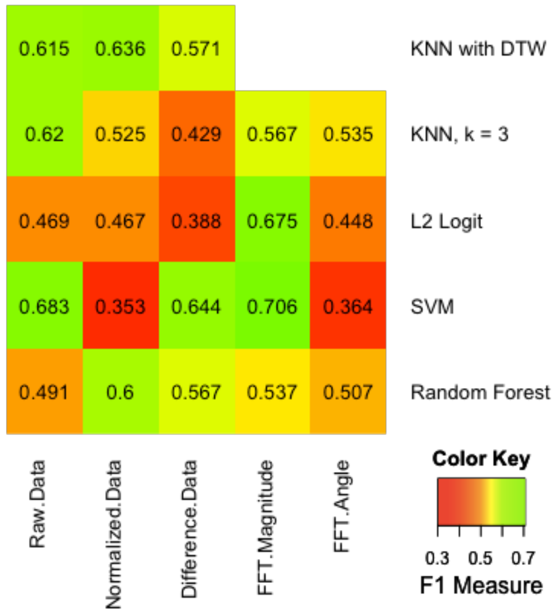

Figure 2. F1 measures comparing supervised methods applied to raw and processed data.

Analyses were preliminarily done using 30-day intervals on all models, yielding the performance results shown in Figure 2. Since the SVM with a Gaussian kernel and KNN with k = 1 using dynamic time warping appear to have the highest F1 measures, more detailed performance statistics are presented in Table 1. The classifer and feature set combination that showed the best performance and nearly the lowest test and training error was KNN with dynamic time warping on zero-mean normalized data, so this method was used to determine the optimal interval length of monitoring for analysis. This classifier and feature set was then applied to varying interval lengths, with performance statistics highlighted in Figure 3.

Discussion

Due to the difficulty seen in previous works attempting to identify trends in AF time series that predict stroke, I attempted an approach combining time series analysis with a wide variety of machine learning classification methods. As seen in Figure 1, the majority of patients have either no or very little AF, and the ones that do have some AF generally have continuous AF, spanning nearly the entire 86,400 seconds per day. Thus, this necessitates identifying patterns from the temporal data in addition to using quantity in order to differentiate cases from controls.

Supervised Learning

Specifically, I chose supervised learning techniques that spanned a wide array of methodologies, from generalized linear models to ensemble methods. Logistic regression was chosen due to its frequent use in clinical data and its interpretability. L2 regularization was applied to avoid overfitting, which was seen in initial analysis of the data. KNN was chosen since it is nonparametric, and previous work has shown that KNN with k = 1 using dynamic time warping has been especially effective for time series classification.[8] SVMs with a Gaussian kernel were used to explore the feature space in a higher dimension, and random forest was used

Table 1. Precision, recall, specificity, and training and test errors for the methods with the highest F1 measures.

KNN, k = 1, Dynamic Time Warping	Precision	Recall	Specificity	Training Error	Test Error
Raw Data	0.621	0.563	0.656	0.344	0.39
Normalized Data	0.618	0.656	0.594	0.375	0.38
Difference Data	0.581	0.563	0.594	0.391	0.42
SVM with Gaussian Kernel	**Precision**	**Recall**	**Specificity**	**Training Error**	**Test Error**
Raw Data	0.560	0.875	0.313	0.387	0.41
Difference Data	0.500	0.906	0.094	0.368	0.38
FFT Magnitude	0.566	0.938	0.281	0.289	0.39

since it is an ensemble method, unlike the others. Overfitting with random forest was seen with a tree number as small as 5, which resulted in a training error of 0, and thus 2 trees were chosen for the analysis.

Model Selection

Though it appears that an SVM with a Gaussian kernel has the highest F1 measures (Figure 1), the specificity is quite low (Table 1). Given that the data is half cases and half controls, this suggests that the model has a tendency to misclassify test examples in a manner that results in a low number of true negatives. KNN with k = 1 using dynamic time warping, on the other hand, appears to perform relatively well on all the feature sets and across all the performance metrics. KNN using dynamic time warping was not applied to Fourier transform data since the frequency data is no longer a time series, so dynamic time warping would not apply as an appropriate distance metric. KNN with dynamic time warping also showed the smallest discrepancy between the training and test errors, as well as having one of the two lowest test errors out of all 23 models, tying with the SVM applied to difference data. In selecting the feature set to proceed with, zero-mean normalized data not only performed the best with KNN with dynamic time warping, but was also an intuitive choice given that the the difference in magnitude between 80,000 and 0 will have an undesired effect on the distance metric when trying to identify fluctuations in AF burden.

Interval Optimization

With up to years of data for some patients, next I was interested in seeing if there was a certain time interval of monitoring before stroke that would be optimal and therefore most sensitive to changes in clinical management. Thus, I applied my chosen model of KNN with dynamic time warping to varying interval lengths. Using the model on 15, 30, 60, 90, and 120 days worth of data, it appears that the optimal interval of these is 60 days. The model applied to 60-day intervals has the highest F1 measure and the lowest training and test error (Figure 3). In fact, the training and test error decreases from 15 to 60 days, and then increases and diverges when using more than 60 days, suggesting that 60 days is the most appropriate length of the feature set. This suggests that there may be some benefit to adjusting anticoagulation based on the previous two months of AF burden.

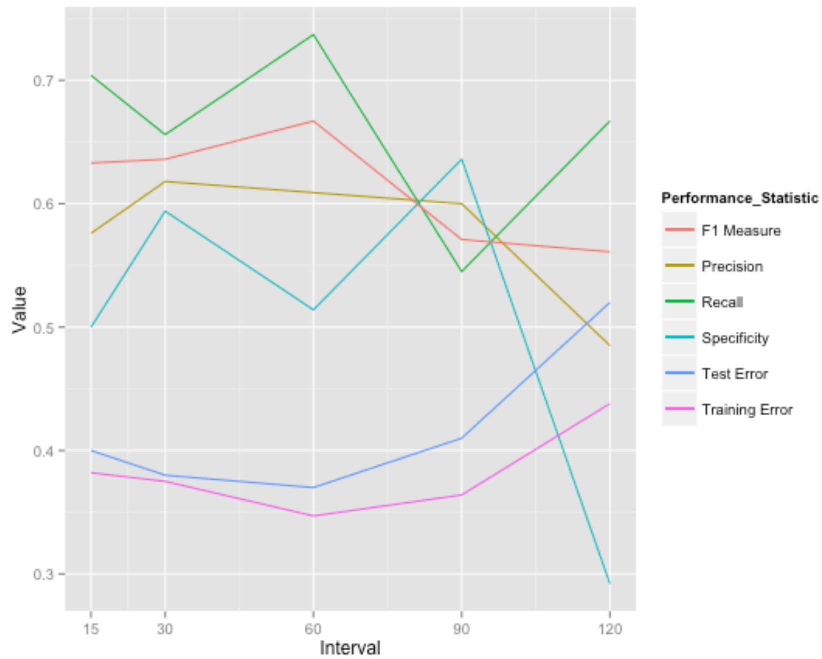

Figure 3. Performance results of KNN with k =1 using dynamic time warping on zero-mean normalized data at varying interval lengths

Conclusions

Using time series atrial fibrillation data to identify patterns and predict strokes has been a difficult problem that has not been met with much success using traditional biostatistical methods. Using machine learning with time series techniques, the highest F1 measure is slightly under 0.7, indicating that identifying patterns still remains a challenge in atrial fibrillation data. Linear and certain non-parametric models were prone to overfitting, even with the use of regularization and limitation of the interval length to trim the feature set. However, by applying a variety of models to explore linear and non-linear decision boundaries and drawing on expertise from the field of time series analysis, the performance of the nearest neighbor using dynamic time warping produced consistently promising results over a range of interval lengths and types of processed data. Further investigation is warranted to explore the relationship between stroke and non-stroke atrial fibrillation patients over time.

Future Directions

Since intervals in controls were selected as the first 60 days of monitoring, it would be interesting to test to see if the results are sensitive to the timing of the control interval. To accomplish this, I would take random 60-day intervals from the total monitoring period in non-stroke patients to serve as the control data. In addition, in cases that have more than 120 days worth of data, doing a case crossover analysis, whereby the first 60 days serves as the control for the last 60 days before stroke, may provide additional insight into this problem. This could be particularly effective given that there may be multiple different types of patterns in cases and in controls, making the groups challenging to model as a whole and thus making it easier to compare cases against themselves.

Acknowledgements

Thank you to the Palo Alto Veterans Administration for providing the data; Susan Schmitt from the Turakhia group for formatting and providing the data; and Mintu Turakhia, Russ Altman, and the CS229 teaching assistants for providing advice and feedback. Additionally, thank you to the rest of the Turakhia research group and the Helix lab. This project was funded by the Dr. Jeffrey Thomas Stroke Shield Foundation.

References

[1] C. T. January, L. S. Wann, J. S. Alpert, H. Calkins, J. C. Cleveland, J. E. Cigarroa, J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. T. Murray, R. L. Sacco, W. G. Stevenson, P. J. Tchou, C. M. Tracy, and C. W. Yancy, "2014 AHA/ACC/HRS Guideline for the Management of Patients With Atrial Fibrillation: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society.," *Circulation*, Apr. 2014.

[2] J. S. Healey, S. J. Connolly, M. R. Gold, C. W. Israel, I. C. Van Gelder, A. Capucci, C. P. Lau, E. Fain, S. Yang, C. Bailleul, C. A. Morillo, M. Carlson, E. Themeles, E. S. Kaufman, S. H. Hohnloser, ASSERT Investigators, "Subclinical atrial fibrillation and the risk of stroke.," *N Engl J Med*, vol. 366, no. 2, pp. 120–129, Jan. 2012.

[3] W. N. Venables and B. D. Ripley. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0, 2002.

[4] T. Giorgino. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, vol. 31, no. 7, pp. 1-24, 2009.

[5] T. Helleputte. LiblineaR: Linear Predictive Models Based On The Liblinear C/C++ Library. R package version 1.80-10, 2014.

[6] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4, 2014.

[7] A. Liaw and M. Wiener. Classification and Regression by randomForest. R News 2(3), 18-22, 2002.

[8] X. Xi, E. Keogh, C. Shelton, and L. Wei. "Fast Time Series Classification Using Numerosity Reduction," Proc. Of the 23rd International Conference on Machine Learning, 2006.