CS229 Project: A Machine Learning Approach to Stroke Risk Prediction

Yu Cao YUFCAO@STANFORD.EDU
Hsu-Kuang Chiu HKCHIU@STANFORD.EDU
Aditya Khosla ADITYA86@STANFORD.EDU
Cliff Chiung Yu Lin CHIUNGYU@STANFORD.EDU

Abstract

In this paper, we consider the prediction of stroke using the Cardiovascular Health Study (CHS) dataset. Missing data imputation, feature selection and feature aggregation were conducted before training and making prediction with the dataset. Then we used a mixture of support vector machine (SVM) and stratified Cox proportional hazards model to predict the occurrence of stroke. Different methods and variations of models were evaluated and compared to find the best algorithm.

1. INTRODUCTION

Currently home-monitored chronic health condition data is used in health-risk assessment systems. According to the data, the systems make predictions on the possibility that a a patient might need medical help in the next few years due to the onset of disease or other conditions. The prediction result is helpful for prevention or early treatment. However, many current systems use relatively simple hand-coded rules to build the prediction models. Applying machine learning techniques to the health-risk assessment problem will be a possible approach to have more accurate predictions. In this project, our objective is to improve the accuracy of stroke prediction using the CHS dataset. It is a challenging task for three main reasons: (1) The CHS dataset suffers from problems such as a large fraction of missing data (25%), sets of correlated features, while some of the features are not highly related to stroke, and (2) the stroke prediction problem itself involves the survival time of individuals that are not captured well by typical machine learning methods (3) the dataset is extremely skewed as only 5% of the patients had a stroke over the period of consideration. Therefore, missing data imputation, feature selection and aggregation, and the Cox proportional hazard models are integrated to overcome problems (1) and (2). To overcome problem (3), we use the area under the receiver operating characteristic (ROC) curve instead of the usual measures of accuracy as we can achieve 95% accuracy simply by classifying all the patients as not having a stroke, which is clearly not a useful prediction.

1.1. The Cardiovascular Health Study (CHS) Dataset

The CHS [8] is a study of risk factors for cardiovascular diseases in people above 65 years old. More than 5,000 patients were examined yearly from 1989 to 1999, with about 2,700 attributes collected annually through medical tests and a set of questionnaires. Events such as stroke and hospitalization were also recorded for each patient. However, in a longitudinal study like CHS, it is unlikely that all patients return each year and provide data for all the required attributes. In the CHS dataset, some attributes are missing intermittently for certain years, while some attributes have missing data for a contiguous series of years. Furthermore, the attributes collected per patient change from year to year and there is no easy way to determine the change over time for a single feature.

1.2. Our Contributions

We compared more than 8 different data imputation methods to find the best method for the given dataset. We also used forward search feature selection to pick a smaller set with 132 features. A prediction model using SVM and Cox is then built to determine whether a patient has a high risk of stroke in the next 5 years. We tried a variety of other methods including EM based algorithms and Gaussian Process regression, but we have not described those as they did not produce sat-

isfactory results. Using data imputation, combined with feature selection, we achieved 0.72 for area under the ROC curve. We also tried to incorporate the data from multiple years to make a better prediction. The rest of the paper is organized as follows. Section 2 provides an overview of the problems that we consider, and reviews the related previous works in the literature. Section 3 describes the approaches to tackle the issues and improve the results. Then, experimental results are shown and discussed in Section 4. Finally, conclusions are in Section 5 and future research directions are given in Section 6.

2. PROBLEM OVERVIEW AND PREVIOUS WORK

The problem we consider can be roughly divided into three parts: (1) imputation of the missing entries in the dataset, (2) selection of strong features and aggregation of weak features, and (3) stroke prediction with the selected filled-in features.

2.1. Missing Data Imputation

The first part of the problem is to fill in the missing entries in the dataset. In [1] several methods were discussed, including filling missing features with column mean, column median and hot-deck imputation. They are commonly used in statistics and serve as the baseline methods against which we would like to test the other algorithms that we develop or use.

We evaluated the missing data imputation results with the following two sets of metrics.

1. Data imputation accuracy: The primary target of missing data imputation is to achieve accurate imputation of the missing entries, evaluated by the following 4 metrics as described in [1].
   (a) Root-Mean-Square Deviation (RMSD)
   (b) Mean Absolute Deviation (MAD)
   (c) Bias: the difference between mean of the imputed results and mean of the ground-truth values.
   (d) Proportionate Variance (PV): the proportion of variance of the imputed results to variance of the ground-truth values.
2. Overall stroke prediction quality: The ultimate goal of the missing data imputation is to collaborate with stroke prediction algorithm. Imputation results were fed to the prediction methods to evaluate the overall stroke prediction quality.

2.2. Feature Selection and Aggregation

The CHS dataset has a large number of attributes ranging from demographic information, clinical history, to biomedical and physical measurements [10]. However, only a small subset of attributes is highly relevant to stroke prediction. In addition, some individual attributes can be weak but correlated, and the aggregated feature may serve as a good indicator to stroke occurrence. [3] has shown that SVM is one of the best methods for feature selection. Other papers such as [10] also use manually selected features according to risk factors analyzed by medical and clinical study. The subset of features selected can be combined with stroke prediction models to evaluate the performance of feature selection and aggregation.

2.3. Stroke Prediction

After filling the missing data entries and selecting the most representative features, we can use those preprocessed data to build the stroke prediction model. In [3], several machine learning algorithms were applied in a stroke risk assessment problem: support vector machines (SVM), decision trees, nearest neighbors, and multilayer perception. According to [3], SVM is the most promising algorithm with high sensitivity and specificity. Therefore SVM was chosen to build our stroke risk prediction model. The evaluation metric in medical diagnosis is better chosen as the area under ROC curve in order to assess both the sensitivity and specificity performance of the model.

The Cox proportional hazards model is one of the most important statistical models used in medical research[9]. This model has been extensively studied[6, 9], and has been applied in various medical applications for the prediction of various diseases[10, 5] and analysis of medical data[7]. The paper by Lumley et al(2002), which makes use of the Cox proportional hazards model is the paper that we used as our baseline result as it contained the best stroke prediction scores as compared to the other papers relating to this application. We followed the method described in [10] to attempt to achieve the same results and to use that as our comparison metric to gauge the success of our models.

3. ALGORITHMS

3.1. Missing Data Imputation Methods

For missing data imputation, two main observations on the CHS dataset affect our strategies: (1) only a small subset of the features is highly related to the occurrence of stroke, and (2) many of the features are

discrete instead of continuous. Due to observation (1), we focus our prediction and evaluation on a set of 132 features selected by forward search using SVM. And due to observation (2), for most algorithms that we implement, we also evaluate extra versions that align each imputed value to its closest discrete label.

Besides the baseline methods, the following imputation algorithms were implemented and evaluated.

1. Column mean with alignment to the closest discrete value
2. Linear regression
3. Linear regression with alignment to the closest discrete value
4. Singular Value Decomposition (SVD)
5. Singular Value Thresholding (SVT) from [4]

3.2. Feature Aggregation

In the set of questionnaires answered by each patient, there are several groups of contiguous questions designed to evaluate a similar quality. For example, how often the patient does various sports or whether the patient can spell some words correctly. The answer options also have the same pattern for these questions. These features are highly correlated. Therefore we looked through the descriptions for each feature and decided to aggregate a group of contiguous features if they are targeted at the same type of assessment and share the same set of values in the choice of answer. In the end, 13 groups were selected. Since the features within a group have the same answer values, we could use the mean of these values as an indicator for the aggregated feature. However there can be missing value in any of the features for a patient. Therefore we should check whether missing values of features exist for each patient, and exclude the missing values when computing the mean for that patient. In addition, the aggregated features should be computed before standardizing all the feature values and removal of features with lots of missing data to ensure that the aggregated features capture the property of the original data.

3.3. Feature Selection

Firstly, features with missing data rate higher than a threshold value were removed. Even though we have missing data imputation algorithms, features with too many missing entries still may not give us accurate information after imputation. Therefore, those features were filtered out.

Then forward search technique was applied to select the subset of most representative features. However, the number of features was 800 after preprocessing, so it will be computationally expensive to complete the entire forward search process with all the features. Thus, L1 regularized logistic regression was executed first to choose 200 features with highest significance ranking as the domain of our forward search process. The forward search was performed with linear kernel SVM and 5-fold cross validation. The value of parameter C was also tuned to achieve the best results.

3.4. Stroke Prediction with Support Vector Machine

Our stroke prediction model has five steps:

1. Feature aggregation: average weak but correlated features
2. Data preprocessing: remove features and training examples with missing data rate higher than threshold values.
3. Features selection: select the best subset of features.
4. Missing data imputation: fill in values of missing entries.
5. SVM training and testing: With the feature set obtained from previous step, train an SVM model with linear kernel due to computation efficiency. In testing, we used 10-fold cross validation to obtain an average generalization performance.

3.5. Cox proportional hazards model

The proportional hazards regression model is given by

$$h(t|\mathbf{X}) = h_0(t)exp(\sum_{i=1}^{n} \beta_i X_i) \tag{1}$$

where $h(t|\mathbf{X})$ is the hazard value at time t given the feature set $\mathbf{X}$ for an individual, $X_1, \ldots, X_n$ are the features, $h_0(t)$ is an arbitrary baseline hazard function, and $\beta_1, \ldots, \beta_n$ are the parameters that we are trying to estimate for the model. This model is known as a semiparametric model because the baseline hazard function is treated nonparametrically. Thus, we can see that the parameters have a multiplicative effect on the hazard value which makes it different from the linear regression models and these models have been shown to correspond better to biological data[2].

Given the attributes of two individuals, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, we can obtain their hazards ratio as

$$\frac{h(t|\mathbf{X}^{(1)})}{h(t|\mathbf{X}^{(2)})} = exp(\sum_{i=1}^{n} \beta_i (X_i^{(1)} - X_i^{(2)})) \tag{2}$$

Thus, we observe that the hazard ratio is independent of the time t, and by comparing these hazard ratios, we can find a threshold to classify the individuals.

The same steps as described in the previous section were used with this model to make predictions to be able to compare the two models.

3.6. Cox proportional hazards model with SVM

The SVM and Cox models described in the previous sections only allowed us to make predictions using the baseline data, or data from a single year. To incorporate the effect of the data from multiple years, the two models were combined using the following algorithm. For each of the years that we want to consider:

1. Perform feature selection using forward search
2. Get a hazards value by applying the Cox proportional hazards model
3. Combine the hazards value by using them as features for a SVM

This model allowed us to incorporate the fact that the features from each year affected the prediction score in a multiplicative way, and the relative importance of each year could be estimated by finding the parameters corresponding to the hazards ratio from each year. The regular model that allows for time dependence in the Cox model could not be used as features do not remain consistent from year to year. This model allowed us to incorporate any set of features from multiple years.

4. EXPERIMENTAL ESTIMATION

4.1. Missing Data Imputation Quality

The missing data imputation results with baseline methods and our proposed algorithms are shown in Table 1, 5 and 6 respectively. In general, linear regression achieves the least RMSD and MAD values with reasonable bias and PV values, and alignment to discrete values further lowers the MAD value but increases RMSD and Bias values.

4.2. Feature Selection and Feature Aggregation Results

Using forward search and 5-fold cross validation with linear kernel SVM, the area under ROC curve achieves the largest value 0.759 when 132 features are selected. The properties of the 13 aggregated features are summarized in Table 7.

4.3. Stroke Prediction Precision with SVM

The average performance of stroke prediction models with SVM and different feature sets is shown in Table 3. Here column mean was used for missing data imputation. We chose a linear kernel with a fixed C value equal to 0.0003 for SVM model. We can see that using only the feature set by forward search has better performance than using manually selected features [10] and using all CHS features. Forward search technique has successfully selected the most representative features in the stroke risk prediction problem. Moreover, we also tried the feature set obtained by forward search beginning with Lumley's features [10]. The result is slightly lower than pure forward search. Feature aggregation combined with the 132 selected features yields similar performance as well.

The various methods of data imputation were combined with the SVM model for verifying the importance of good data imputation. The list of data imputation methods, and their results are listed in Table 2. We found that the different methods of data imputation did not affect the outcome of the model significantly. Furthermore, the RMSD values obtained for the various data imputation methods were not exactly correlated to the stroke prediction outcome. Instead, we found that the data imputation models that gave us the lowest RMSD values such as linear regression and linear regression with discretization produced worse results for area under the ROC curve than simple imputation methods such as discretized column mean and column median. The reason for this may be the fact that the initial feature selection algorithm was run using data that used column mean as the method for data imputation.

4.4. Cox Proportional Hazards Model

We implemented the Cox proportional hazards regression model using the same set of features as described in [10], and found the coefficients by fitting the data using Matlab in a similar fashion as described in the paper. However, we were not able to obtain the same coefficients as described in the paper, and found that the area under the ROC curve was only 0.7021 using the features given in the paper, and 0.7064 using the 132 features found using forward search.

Furthermore, to try to emulate the [10], we followed the method for processing the data exactly as described in the paper and tried to use the same coefficients for the generated data, and found a an area under the ROC curve of 0.5681. This was very low as compared to the reported area of 0.73 in the paper. We were not able to verify this value despite following through the details provided in the paper multiple times. Also, we requested Professor Lumley for his data set to be able to verify his result but we have not

heard back from him for the last 5 weeks.

4.5. SVM combined with the Cox Model

The results of the model combined for two years is given in Table 6. The results from using data for multiple years proved to be the best as we managed to achieve area under the ROC curve values close to 0.73. The area under the ROC curve was averaged over 10 independent training and test data sets. This model was only tried for data from two years due to the lack of time and is possibly a promising direction to achieve better overall results by combining the data in this way for more years.

5. CONCLUSIONS

This project has shown that machine learning algorithms can be a powerful tool to achieve good results even in difficult data sets like the one presented in this paper. We managed to match the values achieved by hand selected features through automatic feature selection. Furthermore, we found that the machine learning tools provide a strong mechanism to handle a variety of tasks involving both imputation of missing data and stroke classification.

6. FUTURE WORK

Due to the shortage of time, we were unable to implement some of the suggestions that were brought up during the course of the project. We would like to suggest these as possible directions for future work. Firstly, for data imputation, an ensemble of methods could be tried that encompass various machine learning methods that are selected from a variety of methods such as SVM, linear regression, logistic regression, EM based methods, etc. Each of the methods could be tested for each feature, and depending on the results, we could use the best method found for each of the features.

Furthermore, upon looking more closely at the data, we found that the data contained values for certain features that could lead to poor classification as well as data imputation. For some of the features that were 'Yes/No' questions, these answers were assigned a value of 1/0, but sometimes there was another value 'Unknown' which was written in the data as 9. Removing these values and applying better data imputation techniques could lead to an overall increase in the stroke prediction quality although the initial tests did not suggest that data imputation improved prediction scores.

Secondly, for stroke prediction, we could try increasing the number of years considered for the Cox-SVM model to potentially increase the prediction score. Also, we could attempt using other methods for exploiting the time series property of the data which was not completely used in our current project. This time-series property could also be used to improve data imputation, but this problem is a difficult one given the nature of the data. The features in the data vary from year to year and even the same features do not have the same names. Thus, it would be a cumbersome task to find commonality between features from multiple years.

ACKNOWLEDGEMENTS

We thank Honglak Lee for his guidance and support throughout the project. He provided us with direction and various useful tools to carry out the data analysis that we needed. Also, we thank Professor Ng for giving us an opportunity to work on such interesting topics by offering this class and for his invaluable teachings throughout the quarter. Lastly, we thank Anand Iyer for partaking in discussions about the project during its inception.

References

[1] J. M. Engels and P. Diehr. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968 976, 2003.

[2] Klein J. and Moeschberger M. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, 2003.

[3] J.C. Sanchez J. Prados, A. Kalousis. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, 4:2320–2332, 2004.

[4] E. J. Candes J.F. Cai and Z. Shen. A singular value thresholding algorithm for matrix completion. *arXiv*, 2008.

[5] Satoshi Saitoh Kenji Ikeda, Hiromitsu Kumada. Effect of repeated transcatheter arterial embolization on the survival time in patients with hepatocellular carcinoma. *Cancer*, 2006.

[6] Tsuyoshi Nakamura Kouhei Akazawa. Simulation program for estimating statistical power of cox's proportional hazards model assuming no specific distribution for the survival time. *Elseview Ireland*, 1991.

[7] Steven G. Self Kung-Yee Liang and Xinhua Liu. The cox proportional hazards model with change point: An epidemiologic application. *Biometrics*,

46:783–793, 1990.

[8] P. Enright L.P. Fried, N.O. Borhani. The cardiovascular health study: design and rationale. *Ann Epidemiol*, 1(3):263–276, 1991.

[9] Thomas Augustin Ralf Bender and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24:1713–1723, 2005.

[10] R. A. Kronmal T. Lumley. A stroke prediction score in the elderly: validation and web-based application. *Journal of Clinical Epidemiology*, 55(2):129–136, 2002.

Model	AUC_ROC^
Cox model(year 1 with feature selection)	0.7064
Cox model(year 2 without feature selection)	0.6403
Cox-SVM model on above data sets	0.7240
Cox-SVM model with feature selection for both years	0.7296

Table 4. Cox proportional hazards model with SVM

	Column Mean			Column Median			Hot Deck		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
Training RMSD	0.4997	0.2259	0.04	0.695	0.2479	0.0406	0.7022	0.3152	0.0548
Test RMSD	0.5002	0.2265	0.0033	0.705	0.2487	0	0.7069	0.3158	0.0004
Training MAD	0.4994	0.1631	0.0061	0.483	0.1302	0.003	0.4975	0.1838	0.0061
Test MAD	0.4999	0.1641	0.0033	0.497	0.1317	0	0.5025	0.1854	0.0001
Training Bias	0.0008	0	-0.0013	0.471	0.0399	-0.483	0.0135	0	-0.0128
Test Bias	0.0406	-0.0005	-0.0454	0.4851	0.0394	-0.497	0.042	-0.0009	-0.0599
Training PV	0.2311	0.0084	0	0.0814	0.0054	0	1.172	0.9445	0.3375
Test PV	0.0116	0.0002	0	1	0.0229	0	7460	57.5	0.2761

Table 1. Data imputation quality with baseline methods

Algorithm	AUC_ROC^
Hot deck	0.7165
Column mean	0.7113
Column mean discretized	0.7199
Column media	0.7188
Linear regression	0.7141
Linear regression discretized	0.7159
SVD	0.6995

Table 2. Stroke prediction quality for a variety of data imputation methods.

	All (800)*	Lumley's [10]	Forward search	Forward search begin with Lumley's [10]	Forward search with aggregated features
Training Accuracy	0.5593	0.5701	0.5707	0.5598	0.5717
Testing Accuracy	0.5465	0.5696	0.5635	0.5557	0.5673
AUC_ROC^	**0.6478**	**0.6306**	**0.6998**	**0.6887**	**0.6986**

Table 3. Results of stroke prediction models with SVM using different feature sets

*The number in parenthesis is the size of feature set.

^AUC_ROC: area under ROC curve

	Column Mean and Alignment			Linear Regression			Linear Regression and Alignment		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
Training RMSD	0.695	0.2443	0.04	0.3752	0.1289	0	0.4333	0.1354	0
Test RMSD	0.705	0.2449	0	0.6998	0.1762	0	0.5843	0.1758	0
Training MAD	0.483	0.1342	0.003	0.317	0.0926	2.21E-07	0.2552	0.0571	0
Test MAD	0.497	0.1355	0	0.4395	0.1212	2.93E-07	0.3414	0.0836	0
Training Bias	0.471	0.0298	-0.483	0.0191	-0.0001	-0.0188	0.1076	0.0061	-0.1273
Test Bias	0.4851	0.0293	-0.497	0.0598	-0.0009	-0.0837	0.1222	0.0051	-0.1689
Training PV	0.152	0.0054	0	1	0.5866	0.1895	1.0177	0.6432	0
Test PV	1	0.0224	0	5.08	0.0541	0.0001	1.7028	0.6571	0.0001

Table 5. Data imputation quality with non-baseline algorithms

	Fill-In with SVD			Fill-In with SVT		
	Max.	Avg.	Min.	Max.	Avg.	Min.
Training RMSD	7.0258	1.4448	0	0.9984	0.4229	0.0541
Test RMSD	6.8523	1.4334	0	1	0.4223	0
Training MAD	5.2127	1.0635	0	0.9967	0.3181	0.003
Test MAD	5.2009	1.0518	0	1	0.3176	0
Training Bias	0.1342	-0.0005	-0.1432	0.9967	0.3181	0.003
Test Bias	0.278	0.005	-0.3097	1	0.3176	0
Training PV	198.543	41.398	1	0	0	0
Test PV	2.43E+05	2995	0.9993	0	0	0

Table 6. Data imputation quality with other non-baseline algorithms

1	Ability to walk without difficulty	2	Optimistic or pessimistic
3	Recent exercise or physical work	4	Difference in physical stamina
5	General trend of mood	6	Sleep problems
7	Awareness of time and venue	8	Simple mathematical ability
9	Ability to spell simple words	10	Ability to repeat words
11	Perform simple tasks with hands	12	Frequency of taking fruits/fruit juices
13	Frequency of eating various beans		

Table 7. Properties of aggregated features