Black Box Stochastic Process for Semi-Supervised Feature Selection

Xochi Kuo
xwatts@alumni.stanford.edu

Abstract

Efficient feature selection is critical for classification tasks, reduces the cost to acquire labels in semi-supervised learning (SSL), and allows for salient features to be identified more efficiently. This work proposes a feature selection wrapper method based on a black box stochastic process of the classification model. The black box stochastic process utilizes all classification model input and output variables to conduct feature selection. The black box stochastic process allows feature selection methods to take into account supervised and unsupervised data without the need for creating pseudo-labels by introducing a technique called censoring. Empirically, the black box method receives an average of 6% greater accuracy than current state-of-the-art semi-supervised stochastic feature selection methods using several real-world datasets, an average 99.35% decrease in running time, and theoretically shows an improvement in the asymptotic behavior of the time complexity of SSL stochastic feature selection. As a result of using a novel stochastic process, the work introduces new metrics to classification models and contains a proof that defines the probability distribution of recall, which is a well-known statistic in classification.

1 Introduction

Feature selection is an important dimension reduction technique. It identifies predictive features within noisy data, excludes unrelated features, reduces overfitting in predictive models, provides better generalization of predictive models, and reduces variance in predictive models. [Hastie *et al.*, 2017] In semi-supervised learning (SSL), feature selection methods have large proportions of missing or unknown target information. Labels for classification are not always available or are costly to acquire. Therefore, many fields such as medicine [Liu *et al.*, 2020] rely on semi-supervised feature selection including information security [Gu *et al.*, 2019] and computer vision [Zhao *et al.*, 2008]. Certain fields such as robotics, biological health, and civil infrastructure, specifically rely on stochastic semi-supervised feature selection. [Sechidis *et al.*, 2018; Fei *et al.*, 2021; Li *et al.*, 2019]

Semi-supervised feature selection methods introduce pseudo-labels for unlabeled data. Pseudo-labels are prone to introduce error and may become difficult and hazardous. The proposed black box method uses a technique called censoring on unlabeled observations to conduct feature selection without needing to predict the labels of unlabeled data. Censoring is performed in the statistical field of survival analysis. The black box stochastic process presented in this paper describes the classification model as a stochastic process modeling the features, known and unknown labels, and classification model score. Treating the classification model as a black box allows both labeled and unlabeled data to train the feature selection model without needing to generate pseudo-labels.

The main contributions of the work are summarized below.

- The innovative application of a black box stochastic process which characterizes a classification model represents all variables in the modeling process - the feature data, labels, and classification model score. Censoring, a new technique in SSL, is introduced and avoids the need for pseudo-labels.
- The black box method using penalized Cox Proportional Hazards Model for feature selection boosts the accuracy of current semi-supervised stochastic feature selection methods, notably having comparable performance to the original features, and improves the experimental running time and theoretical upper bound to $O(np)$ for current semi-supervised stochastic feature selection.
- Recall, a well-known classification model metric, is proven to equal the product limit estimator of the black box stochastic process. Recall at $R(s_j)$ is defined as follows where d_i is an indicator for positive observations $1...k$ and the classification scores $s_1, ..., s_{k-1}, s_k$ such that $s_1 < ... < s_{k-1} < s_k$.

$$R(s_j) = \frac{\sum_{s_i > s_j} d_i}{\sum_{s_1 \leq s_j \leq s_k} d_i} \tag{1}$$

2 Related Works

Semi-supervised feature selection methods are divided into filter, wrapper, and embedded methods.[Sechidis and Brown,

2018] Many semi-supervised wrapper methods generate pseudo-labels for feature selection [Lin *et al.*, 2020; Jiang *et al.*, 2020; Zhang *et al.*, 2019; Sechidis *et al.*, 2018; Sechidis and Brown, 2018]. Incorporation of labeled data is expected to provide a higher discrimination power compared to unsupervised methods. The effectiveness of the classifier to generate pseudo-labels depends on the accuracy of the features leading to a negative feedback loop and ignore the intra-class variations of the feature space.[Ren *et al.*, 2008; Zeng *et al.*, 2016]. Although pseudo-labels can achieve good performance in many applications, they are sensitive to the bias that labeled data and unlabeled data share information, whether local discriminative information or global discriminative information, so are subjective to noise.

A combination of unsupervised and supervised feature selection has been applied to semi-supervised methods to avoid pseudo-labels. The first method uses unlabeled observations to discover the structure of the data and labeled observations to select features that preserve the within-class graph and between-class graph structures.[Zhao *et al.*, 2008] Another method isolates the positive class set of the labeled data to maximize the feature selection optimization and demonstrates good performance on noisy data.[Jensen *et al.*, 2015]

Stochastic processes in feature selection are found in semi-supervised and unsupervised data. Unsupervised approaches disregard all known labels where Markov transition probabilities are used to obtain the data structure and describe the relationship between adjacent points and points farther away in order to conduct feature selection. [Min *et al.*, 2021] They disregard label information in feature selection and the noise introduced from the local data structures. The stochastic feature selection with the most impact in performance and efficiency in semi-supervised data is a simple yet effective strategy which assigns all unlabeled data with positive pseudo-labels or all unlabeled data with negative pseudo-labels, then applies known feature selection methods - such as Mutual Information Maximization (MIM), Joint Mutual Information (JMI), and Incremental Association Markov Blanket (IAMB) - Semi-JMI, Semi-MIM, and Semi-IAMB feature selection. [Sechidis and Brown, 2018; Fei *et al.*, 2021] This method performs best when the labels are missing (not at random) or there is prior knowledge of the distribution of unlabeled observations. Fuzzy c-means clustering unlabeled examples improves the performance of the Semi-JMI algorithm. [Li *et al.*, 2019] A limitation of the work, however, is that mutual information is computed between binary categorical data, so all continuous features are forced to categorical values. The authors received greater performance in higher order Conditional Mutual Information (CMI) and higher order JMI. [Sechidis *et al.*, 2018] These computations trade off computational complexity to calculate higher-order information for predictive feature selection and greatly diminish the efficiency of the method.

3 Background

The black box method introduced in this paper is a wrapper stochastic process using methods from survival analysis that circumvent pseudo-labels in semi-supervised feature selection. The black box method is the first time a counting process of the inputs and outputs of the classification model has been used for the purpose of feature selection.

Survival analysis [Aalen *et al.*, 2008] is a special branch of statistics that aims to measure the duration to which an event will take place. [Spooner *et al.*, 2020] Typically survival analysis models the time to event, however the methods may be used on an index other than time [Aalen *et al.*, 2008]. Survival analysis data consists of three types of variables. These variables are the index variable, the event variable, and covariate variables (or features) of the event. In contrast, machine learning classification models only define two variables; the features and labels. Survival data is high-dimensional, heterogenous, and censored, so special techniques for feature selection in survival models have been created specific to this type of data.[Spooner *et al.*, 2020]

Survival analysis is a stochastic Markov process, which has the characteristic Markov property or memoryless property. [Paul and Baschnagel, 2013] The memoryless property states that once the current state of the process is known, any knowledge of the past (or any circumstance in which an observation receives a lesser score) does not give any further information about the state of the process in the future (or in this case a higher score). It suffices to make use of the current state of the covariates (features) and event (label) to describe the probability distribution of the process over the score interval [Paul and Baschnagel, 2013].

After the score to event process is described, the well-known survival functions can be applied to the process. The functions and metrics commonly used in survival analysis are the survival function and the hazard function. The survival function represents the probability that an individual survives from the time of origin to some time beyond time t. It is usually estimated by the product limit estimator, also known as the Kaplan Meier curve.[Kaplan and Meier, 1958] The product limit estimator is the maximum likelihood estimator of the cumulative distribution function (CDF) [Kaplan and Meier, 1958] of the probability of event when no observations are censored. The second function in survival analysis, the hazard function, gives the instantaneous potential of having an event at a time point, given survival up to that time. The strength of the relationship between the survival time and the observation-related features or covariates is determined using the hazard function in multiple regression models. Current feature selection methods use the coefficient values from these learning algorithms to select features in survival data.[Pölsterl, 2020; Neums *et al.*, 2019; Wang *et al.*, 2019]

Feature selection models include penalized Cox's proportional hazards regression models, comprising of LASSO penalized model [Simon *et al.*, 2011] with L_1 regularization and elastic net penalized model [Simon *et al.*, 2011] with L_1 and L_2 regularization terms. [Pölsterl, 2020] The penalized Cox proportional hazards model is described below where the hazard function is $\alpha(s)$, the baseline hazard is $\alpha_0(s)$, the regression parameters are $(\beta_1, ..., \beta_p)$, and the p covariate features are $(x_1, ..., x_p)$.

$$\alpha(s|\mathbf{X}) = \alpha_0(s)\exp(\sum_{m=1}^{p}\beta_m X_m) \tag{2}$$

The penalized log partial likelihood $\ell(\beta)$ is the following where o_p and r_p correspond to tuning parameters. The parameters $o_p \geq 0$ controls the amount of shrinkage and $r_p \in [0,1]$ is the relative weight of the L_1 and L_2 penalty.

$$\log \ell(\beta) - o_p(r_p \sum_{m=1}^{p}(|\beta_m|) + \frac{1-r_p}{2}\sum_{m=1}^{p}(\beta_m^2)) \tag{3}$$

Cox's proportional hazards (CPH) model is a common linear model used in survival analysis, with the assumption that the covariates are multiplicatively related to the hazard.[Cox, 1972] Penalized CPH models implement regularization and maximize the partial likelihood of the coefficients by shrinking the value of the coefficients. Adding the LASSO penalty allows the model to select a subset of features that are predictive. [Neums *et al.*, 2019; Wang *et al.*, 2019]

Elastic net CPH model uses a weighted combination of the Ridge L_2 and LASSO L_1 penalties. Elastic net is able to select features to set to zero as well as have a solution for features that are highly correlated or complementary by shrinking the coefficients rather than eliminating them completely. [Zou and Hastie, 2005; Simon *et al.*, 2011]

The remainder of the paper defends the implementation of survival analysis, an application of the black box stochastic process, to the classification model input and output data. State-of-the-art survival feature selection techniques can improve the classification model once the black box method is assumed.

4 Method

The black box method models the causal effect of the feature data on the classification score of the positive class labels. After formulation of the stochastic process, a feature selection model can then be utilized to rank the features and select the most important features.

Theorem 1 (Classification Model Stochastic Process)**.** *The stochastic process event is an observation changing from a negative class observation to a positive class observation over the indexed value of the model score and the feature data are covariate data for the observation.*

In the black box stochastic process, the event of interest is a positive class label. The negative class observations can be truncated because there is zero probability to go from a negative class observation to a positive class observation over the classification model score given any feature set. Truncation is the removal of observations from the analysis and is used when the event times of observations are not detectable due to limitations that exclude it from being part of the stochastic process. The classification model score is an ordered sequence and can summarize and rank specific observations, therefore it can be considered an index and is analogous to time. The feature data are the covariates of the event as they are the features associated with the positive class label.

Algorithm 1 Black Box Stochastic Process Feature Selection

Input: Covariates $x \in \mathbb{R}^m$ for positive and unlabeled observations, $y_{structured} = \{(z_1, s_1), ..., (z_k, s_k)\}$**where**
$z_i = \begin{cases} 1, & \text{if } y_i \text{ is positive and labeled} \\ 0, & \text{if } y_i \text{ is censored/unlabeled} \end{cases}$
s_i is the classification model score for positive and unlabeled observations
Parameter: o_p, r_p
Output: selected f features

1: Solve for $\hat{\beta}$.

$$\alpha(s|\mathbf{X}) = \alpha_0(s)\exp(\sum_{m=1}^{p}\beta_m X_m) \tag{4}$$

by maximizing the $\log \ell(\beta)$

$$\log \ell(\beta) - o_p(r_p \sum_{m=1}^{p}(|\beta_m|) + \frac{1-r_p}{2}\sum_{m=1}^{p}(\beta_m^2)) \tag{5}$$

using the package `scikit-survival` [Pölsterl, 2020]

2: Sort $|\hat{\beta_m}|$ in descending order where $m \in 1, ..., p$.
3: **Return** the first f features in m.

To proceed, the reader must be convinced that the stochasticity of the black box model lies in the observation's feature data, and that feature data causes the positive class label to occur at a score point. For example, consider a classification model which predicts whether a person will purchase dog food with features such as demographics, number of pets owned, and zip code. Some people will never buy dog food because they are not dog people. This simplified version of the world makes the assumption that all people who don't buy dog food are not dog people and truncates those observations. Additionally, a person's features has randomness, where at one point Bill might live in a tiny apartment then move to a house in the suburbs, however the classification model remains the same but would give Bill different scores depending on his zip code. Assuming Bill owns a dog and is a positive class, the $\hat{\beta}$ values in the feature selection algorithm measures how much the score changes depending on Bill's zip code. For someone who doesn't have a label, there is a possibility that he or she would be dog person and would buy dog food, so these unlabeled observations are censored and included in the black box model.

Theorem 2 (Black Box Markov Stochastic Process)**.** *The black box method is a Markov process because it has the memoryless property.*

The positive class state $Y(s) = 1$ is absorbing. The intensity of leaving the negative class state $Y(s) = 0$ and entering $Y(s) = 1$ is $\alpha(s)$ at score s. The transition probabilities only depend on the score s and not on the starting score, therefore they are score-homogenous and have the Markov property. Let the selection of classification score points be $s_1, ..., s_{k-1}, s_k$ such that $s_1 < ... < s_{k-1} < s_k$.

$$P(Y(s) = y| \\ Y(s_k) = y_k, Y(s_{k-1}) = y_{k-1}, ..., Y(s_1) = y_1) \\ = P(Y(s) = y|Y(s_k) = y_k) \tag{6}$$

The following Markov transition probabilities describe the probability for the process to move from one state to another within a specified score interval.

Definition 1 (Black Box Transition Probabilities)**.** *Let the positive class labels be $y_1, ..., y_{k-1}, y_k$ and the survival function $I(s) = P(S > s)$.*

$$\alpha(\mathbf{s}) = \begin{bmatrix} & \textbf{Y(s) = 0} & \textbf{Y(s) = 1} \\ \textbf{Y(s) = 0} & -\alpha(s) & \alpha(s) \\ \textbf{Y(s) = 1} & 0 & 0 \end{bmatrix} \tag{7}$$

The full probability transition matrix is the following:

$$P(s) = \begin{bmatrix} & \textbf{Y(s) = 0} & \textbf{Y(s) = 1} \\ \textbf{Y(s) = 0} & I(s) = P(S > s) & 1 - I(s) \\ \textbf{Y(s) = 1} & 0 & 1 \end{bmatrix} \tag{8}$$

Corollary 1 (Censor Unlabeled Observations)**.** *An observation with an unknown class label, in the case of semi-supervised models, is censored where the known information about the observation is only the classification score and feature data but not the true class label.*

An observation can be in the positive class, negative class, or unknown. Unlabeled observations still provide some information even without the class labels and could still potentially be positive observations. An observation is censored when the true event and index of an observation is not known at the time of the test or is not measured during the duration of the study. In the example above, a person could still have purchased dog food without the data scientist having data from their favorite grocery store. The labeled observations would include all people where grocery store purchases are available and unlabeled observations where purchases are unavailable. It is important to not confuse censoring and truncation.

Feature selection regression analysis is based on the estimated survival function (also known as the product limit estimator) and the hazard function estimator. The following product limit estimator and hazard function are defined for the stochastic process where d_i is an indicator for the positive observations $1...k$ and m_i is an indicator for positive and unlabeled observations $1...k$.

Definition 2 (Product Limit Estimator)**.**

$$\hat{I}(s_j) = \begin{cases} 1, & \text{if } s_j < s_1 \\ \prod_{s_i \le s_j} [1 - \frac{d_i}{M_i}], & \text{if } s_i \le s_j \end{cases} \tag{9}$$

where

$$M_i = \sum_{s_i \ge s_j} m_i \tag{10}$$

Definition 3 (Cumulative Hazard Function Estimator)**.** *The standard estimate for the cumulative hazard function is defined as follows.*

$$\tilde{A}(s_j) = \begin{cases} 0, & \text{if } s_j \le s_1 \\ \sum_{s_i \le s_j} \frac{d_i}{M_i}, & \text{if } s_i \le s_j \end{cases} \tag{11}$$

The product limit estimator function is actually a popular machine learning metric. The following proof shows that the product limit estimator is equivalent to the recall curve $R(s)$ when there is no censoring. The product limit estimator has a known probability distribution definition. It can then be concluded that $R(s)$ has a probability distribution definition as the probability that a positive class observation receives a score beyond score s under the black box model. Equivalence is established by first taking the logarithm of the product limit estimator, rearranging the terms in the summation, then taking the exponential function of the result to arrive to the function for recall.

Proof. When

$$M_i = \sum_{s_i \ge s_j} m_i = \sum_{s_i \ge s_j} d_i \tag{12}$$

$$\begin{aligned} \log_e \hat{I}(s) &= \begin{cases} \log_e 1, & \text{if } s < s_1 \\ \log_e \prod_{s_i \le s} [1 - \frac{d_i}{M_i}], & \text{if } s_i \le s \end{cases} \\ &= \begin{cases} 0, & \text{if } s < s_1 \\ \sum_{s_i \le s} \log_e [1 - \frac{d_i}{M_i}], & \text{if } s_i \le s \end{cases} \\ &= \begin{cases} 0, & \text{if } s < s_1 \\ \sum_{s_i \le s} \log_e [\frac{M_i - d_i}{M_i}], & \text{if } s_i \le s \end{cases} \\ &= \begin{cases} 0, & \text{if } s < s_1 \\ \sum_{s_i \le s} [\log_e (M_i - d_i) & \\ - \log_e (M_i)], & \text{if } s_i \le s \end{cases} \\ &= \begin{cases} 0, & \text{if } s < s_1 \\ - \log_e (\sum_{s_i \ge s_1} d_i) & \\ + \log_e (\sum_{s_i > s} d_i), & \text{if } s_i \le s \end{cases} \\ &= \begin{cases} 0, & \text{if } s < s_1 \\ \log_e (\frac{\sum_{s_i \ge s} d_i}{\sum_{s_i \ge s_1} d_i}), & \text{if } s_i \le s \end{cases} \end{aligned} \tag{13}$$

Taking the exponential function of the $\log_e \hat{I}(s)$ returns the following.

$$\begin{aligned} e^{\log_e \hat{I}(s)} &= \hat{I}(s) \\ &= \begin{cases} 1, & \text{if } s < s_1 \\ \frac{\sum_{s_i > s} d_i}{\sum_{s_i \ge s_1} d_i}, & \text{if } s_i \le s \end{cases} \\ &= \begin{cases} 1, & \text{if } s < s_1 \\ \frac{\sum_{s_i > s} d_i}{\sum_{s_1 \le s \le s_k} d_i}, & \text{if } s_i \le s \end{cases} \\ &= R(s) \end{aligned} \tag{14}$$

$$\hat{I}(s) \iff R(s) \tag{15}$$

□

Proof that recall is logically equivalent to the survival function results in showing that recall $R(s)$ has a probability distribution definition under the black box model. $R(s)$ is the probability that a positive observation will receive a score greater than s. [Kaplan and Meier, 1958]

Data Name	#Features	#Obs.	#Classes
wine	13	178	3
splice	61	3190	3
sonar	60	208	3
ionosphere	34	351	2
heart-disease	13	270	2
breast-cancer	10	675	2
adult	14	48842	2
abalone	8	4177	29
codon	69	13028	11

Table 1: Data used for experiments.

5 Experiments

Several stochastic feature selection methods were evaluated including CoxLasso, CoxNet, MIM, JMI, CMI-3, and FuzzyJMI. The feature selection methods select features using training data, retrain the classification models only using the selected features on training data, and finally evaluate the reduced classification models on holdout data. The results display an average of the holdout set from 5-fold cross-validation. Feature selection using the black box method is shown to be effective compared to other stochastic methods using 9 UCI datasets. [Mangasarian and Wolberg, 1990; Dua and Graff, 2017; Khomtchouk, 2020; Nakamura *et al.*, 2000] A summary of the datasets used are listed in Table 1. The datasets were drawn from several domains and consist of low-dimensional and high-dimensional data.

A 5-layer feed forward neural network with 4 layers using relu activation and an additional layer using sigmoid activation was used to learn binary classification. Adam optimizer with 0.001 initial learning rate and binary cross-entropy loss over 100 epochs was used to train the feed forward neural network.

In order to select optimal parameters in the feature selection methods, parameter tuning was conducted through RandomizedSearchCV using 2-fold cross validation prior to fitting the final feature selection model. Parameters were selected based on the best concordance index. Parameter tuning was only done for the penalized Cox models, selecting the o_p ratio in Cox Lasso, and selecting the o_p and r_p ratios in Cox Elastic Net.[Pölsterl, 2020] All penalized Cox models used Breslow's method to handle event ties. [Breslow, 1974] All experiments used a random seed.

Feature selection was performed after ranking the features by each method, then evaluating the hold-out performance of a classification model using 75% of features. CoxLasso and CoxNet used the absolute value of the coefficients of features to rank in descending order and select the top 75% of features. JMI, MIM, CMI-3, and FuzzyJMI ranked and returned the top 75% features.

The black box methods outperformed the benchmark methods in 70% and 90% unlabeled observations for splice, sonar, heart-disease, adult, abalone, wine, and ionosphere data in holdout accuracy after retraining the classification model using 75% of the features available. It is out of the scope of the paper to understand why the black box methods did perform as well as the benchmark methods in breast-cancer and codon. Nevertheless, the black box methods still outperformed the holdout classification accuracy without feature selection in breast-cancer data. Meaningful attributes of the data that led to these results deserve to be further investigated in its own work. On average the black box methods received 0.82 ± 0.14 accuracy, outperforming the benchmark methods which have an average 0.77 ± 0.20 accuracy. This results in a 6.4% improvement in accuracy.

It is worth mentioning that the black box methods are able to input both categorical and continuous features, while the benchmark methods are limited to binary categorical variables and require transformations prior to feature selection.[Sechidis *et al.*, 2018; Sechidis and Brown, 2018]

6 Time Analysis

The time complexity for both the penalized CPH models are the same as CPH being $O(np)$ where n is the size of the sample and p is the number of features. [Wang *et al.*, 2019] The benchmark methods have greater time complexity than the black box methods. The upper bound with memoisation for MIM is $O(np^2)$, JMI $O(np^3)$, CMI-3 is $O(np^4)$ when calculating the top p features in data, where n is the size of the sample and p is the number of features. [Sechidis *et al.*, 2018]

To evaluate the efficiency, the average feature selection time over all trials in wall clock seconds was recorded. The experiments ran on a 2.2 GHz Quad-Core Intel Core i7, 16 GB Ram, macOS Big Sur 11.1. The average time was 0.60 $\pm$ 0.89 seconds for CoxLASSO, 2.61 $\pm$ 7.23 seconds for CoxNet, 13.39 $\pm$ 11.74 seconds for MIM, 41.29 $\pm$ 57.94 seconds for JMI, 886.86 $\pm$ 1474.29 seconds for CMI-3, and 42.97 $\pm$ 61.54 seconds for FuzzyJMI. On average the black box methods took 1.60 $\pm$ 5.23 seconds, while the benchmark methods took 246.13 $\pm$ 824.26 seconds. This results in a much faster algorithm taking 0.65% of the prior running time.

7 Discussion

The black box method improved stochastic process feature selection in both model performance and efficiency of the algorithm. Censoring has the impact of using all labeled and unlabeled data in semi-supervised feature selection. The key for designing an effective semi-supervised feature selection algorithm is to develop a framework under which the relevance of a feature can be evaluated by both labeled and unlabeled data in a natural way[Zeng *et al.*, 2016]. This has been done through censoring in the black box stochastic process.

A restriction of black box stochastic processes confines the application to features with a causal relationship to the label. An example would be EEG data when monitoring adaptive brain-machine interface of brain-controlled vehicles, which uses stochastic feature selection. [Fei *et al.*, 2021] Another restriction is the assumption of truncation for negative class observations. Future research may be done to improve the black box model to incorporate a better model of the world.

Future research may be adapted to positive and unlabeled (PU) feature selection, an area of research where only positive observations are labeled but unlabeled data contains both positive and negative observations. [Bekker and Davis, 2020]

Unlabeled %	Data	BlackBoxCoxLasso	BlackBoxCoxNet	FuzzyJMI	JMI	MIM	CMI3
0	splice	0.698 ± 0.085	0.748 ± 0.109	0.868 ± 0.094	0.861 ± 0.118	**0.873** ± 0.095	0.858 ± 0.107
	breast-cancer	**0.948** ± 0.023	0.942 ± 0.043	0.938 ± 0.029	0.936 ± 0.040	0.933 ± 0.042	0.939 ± 0.038
	sonar	**0.678** ± 0.120	0.664 ± 0.093	0.485 ± 0.116	0.446 ± 0.128	0.465 ± 0.146	0.451 ± 0.142
	heart-disease	0.752 ± 0.107	**0.796** ± 0.051	0.763 ± 0.044	0.767 ± 0.048	0.763 ± 0.044	0.778 ± 0.045
	adult	0.807 ± 0.005	**0.817** ± 0.004	0.796 ± 0.003	0.800 ± 0.003	0.797 ± 0.005	0.800 ± 0.003
	codon	0.849 ± 0.312	0.853 ± 0.300	**0.954** ± 0.097	0.953 ± 0.097	0.953 ± 0.097	0.953 ± 0.097
	abalone	0.839 ± 0.008	**0.843** ± 0.004	0.825 ± 0.017	0.824 ± 0.017	0.823 ± 0.020	0.828 ± 0.015
	wine	0.944 ± 0.052	**0.950** ± 0.053	0.842 ± 0.069	0.848 ± 0.090	0.814 ± 0.083	0.848 ± 0.109
	ionosphere	0.918 ± 0.036	**0.929** ± 0.036	0.843 ± 0.083	0.877 ± 0.093	0.843 ± 0.054	0.860 ± 0.066
70	splice	0.679 ± 0.066	**0.887** ± 0.025	0.777 ± 0.156	0.813 ± 0.094	0.827 ± 0.117	0.825 ± 0.122
	breast-cancer	0.942 ± 0.036	0.936 ± 0.038	0.939 ± 0.037	0.945 ± 0.037	**0.953** ± 0.027	0.924 ± 0.056
	sonar	0.634 ± 0.131	**0.635** ± 0.128	0.345 ± 0.184	0.437 ± 0.129	0.293 ± 0.134	0.379 ± 0.189
	heart-disease	0.752 ± 0.082	**0.807** ± 0.058	0.789 ± 0.043	0.748 ± 0.077	0.733 ± 0.096	0.781 ± 0.053
	adult	0.799 ± 0.002	**0.812** ± 0.006	0.792 ± 0.002	0.789 ± 0.005	0.790 ± 0.002	0.791 ± 0.006
	codon	0.833 ± 0.336	0.859 ± 0.275	0.953 ± 0.097	**0.954** ± 0.097	0.952 ± 0.096	0.954 ± 0.097
	abalone	**0.840** ± 0.012	0.835 ± 0.010	0.810 ± 0.019	0.806 ± 0.018	0.813 ± 0.011	0.803 ± 0.023
	wine	0.888 ± 0.068	**0.916** ± 0.071	0.498 ± 0.246	0.713 ± 0.087	0.651 ± 0.253	0.779 ± 0.205
	ionosphere	0.849 ± 0.070	0.841 ± 0.085	**0.875** ± 0.086	0.829 ± 0.082	0.855 ± 0.084	0.841 ± 0.074
90	splice	0.669 ± 0.083	**0.841** ± 0.027	0.682 ± 0.201	0.726 ± 0.152	0.745 ± 0.146	0.717 ± 0.131
	breast-cancer	0.899 ± 0.045	0.938 ± 0.079	**0.941** ± 0.053	0.933 ± 0.028	0.921 ± 0.037	0.938 ± 0.033
	sonar	**0.615** ± 0.159	0.587 ± 0.146	0.281 ± 0.259	0.280 ± 0.297	0.312 ± 0.225	0.361 ± 0.243
	heart-disease	0.704 ± 0.054	**0.770** ± 0.036	0.689 ± 0.027	0.678 ± 0.064	0.704 ± 0.104	0.707 ± 0.038
	adult	0.790 ± 0.010	**0.803** ± 0.012	0.789 ± 0.007	0.788 ± 0.003	0.780 ± 0.005	0.793 ± 0.006
	codon	0.841 ± 0.318	0.832 ± 0.330	0.953 ± 0.097	0.953 ± 0.097	**0.954** ± 0.097	0.954 ± 0.097
	abalone	**0.836** ± 0.013	0.829 ± 0.025	0.795 ± 0.029	0.801 ± 0.036	0.793 ± 0.010	0.784 ± 0.031
	wine	**0.888** ± 0.062	0.839 ± 0.113	0.586 ± 0.229	0.500 ± 0.219	0.392 ± 0.314	0.397 ± 0.397
	ionosphere	0.786 ± 0.100	**0.829** ± 0.045	0.801 ± 0.109	0.767 ± 0.095	0.798 ± 0.122	0.792 ± 0.113

Table 2: Experiment results comparing the average cross-entropy classification accuracy ± one standard deviation over 0% unlabeled observations, 70% unlabeled observations, and 90% unlabeled observations. Results are given using 75% of features after feature selection. The maximum cross-entropy accuracy is in bold by data set and unlabeled percent. Underlined values indicate the holdout accuracy using feature selection is greater than the holdout accuracy without feature selection.

The black box methods can be extended in future research to include bias correction. Depending on the distribution of the unlabeled data, censoring may introduce bias and would need to be corrected. Second, future research may also include the black box method as a method to conduct counterfactual analysis because censoring can process counterfactual examples. Finally, the black box feature selection methods introduced in this paper are linear models. Further research can be conducted on nonlinear black box methods.

8 Conclusion

The paper proposes a novel black box stochastic process in order to conduct feature selection using survival analysis techniques. The black box method outperformed stochastic feature selection methods in experimental results and improves algorithmic convergence from current stochastic semi-supervised feature selection methods. Censoring is presented as a novel method to use information from the classification model score of unknown label observations, while avoiding pseudo-labels in semi-supervised feature selection methods.

Counting processes are well-studied and have many known properties, which allow for such properties to be applied to classification models. In this work one such interesting property is found, where the product limit estimator of the black box stochastic process has logical equivalence to the recall function, a popular metric in machine learning. There may be more interesting properties yet to be discovered.

References

[Aalen *et al.*, 2008] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[Bekker and Davis, 2020] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

[Breslow, 1974] Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

[Cox, 1972] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Fei *et al.*, 2021] Weijie Fei, Luzheng Bi, and Jingwei Zhang. Adaptive brain-machine interface of brain-controlled vehicles using semi-mim and tsvm. In *Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence*, pages 31–35, 2021.

[Gu *et al.*, 2019] Yonghao Gu, Kaiyue Li, Zhenyang Guo, and Yongfei Wang. Semi-supervised k-means ddos detection method using hybrid feature selection algorithm. *IEEE Access*, 7:64351–64365, 2019.

[Hastie *et al.*, 2017] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2017.

[Jensen *et al.*, 2015] Richard Jensen, Sarah Vluymans, Neil Mac Parthaláin, Chris Cornelis, and Yvan Saeys. Semi-supervised fuzzy-rough feature selection. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pages 185–195. Springer, 2015.

[Jiang *et al.*, 2020] Lin Jiang, Guoxian Yu, Maozu Guo, and Jun Wang. Feature selection with missing labels based on label compression and local feature correlation. *Neurocomputing*, 395:95–106, 2020.

[Kaplan and Meier, 1958] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[Khomtchouk, 2020] Bohdan B Khomtchouk. Codon usage bias levels predict taxonomic identity and genetic composition. *bioRxiv*, 2020.

[Li *et al.*, 2019] Juan Li, Cong Wang, Zhihong Qian, and Changgang Lu. Optimal sensor placement for leak localization in water distribution networks based on a novel semi-supervised strategy. *Journal of Process Control*, 82:13–21, 2019.

[Lin *et al.*, 2020] Mingquan Lin, He Cui, Weifu Chen, Arna van Engelen, Marleen de Bruijne, M Reza Azarpazhooh, Seyed Mojtaba Sohrevardi, J David Spence, and Bernard Chiu. Longitudinal assessment of carotid plaque texture in three-dimensional ultrasound images based on semi-supervised graph-based dimensionality reduction and feature selection. *Computers in Biology and Medicine*, 116:103586, 2020.

[Liu *et al.*, 2020] Chia-Hui Liu, Chih-Fong Tsai, Kuen-Liang Sue, and Min-Wei Huang. The feature selection effect on missing value imputation of medical datasets. *Applied Sciences*, 10(7):2344, 2020.

[Mangasarian and Wolberg, 1990] Olvi L Mangasarian and William H Wolberg. Cancer diagnosis via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1990.

[Min *et al.*, 2021] Yan Min, Mao Ye, Liang Tian, Yulin Jian, Ce Zhu, and Shangming Yang. Unsupervised feature selection via multi-step markov probability relationship. *Neurocomputing*, 453:241–253, 2021.

[Nakamura *et al.*, 2000] Yasukazu Nakamura, Takashi Gojobori, and Toshimichi Ikemura. Codon usage tabulated from international dna sequence databases: status for the year 2000. *Nucleic acids research*, 28(1):292–292, 2000.

[Neums *et al.*, 2019] Lisa Neums, Richard Meier, Devin C Koestler, and Jeffrey A Thompson. Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 415–426. World Scientific, 2019.

[Paul and Baschnagel, 2013] Wolfgang Paul and Jörg Baschnagel. *Stochastic processes*, volume 1. Springer, 2013.

[Pölsterl, 2020] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

[Ren *et al.*, 2008] Jiangtao Ren, Zhengyuan Qiu, Wei Fan, Hong Cheng, and S Yu Philip. Forward semi-supervised feature selection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 970–976. Springer, 2008.

[Sechidis and Brown, 2018] Konstantinos Sechidis and Gavin Brown. Simple strategies for semi-supervised feature selection. *Machine Learning*, 107(2):357–395, 2018.

[Sechidis *et al.*, 2018] Konstantinos Sechidis, Konstantinos Papangelou, Paul D Metcalfe, David Svensson, James Weatherall, and Gavin Brown. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376, 2018.

[Simon *et al.*, 2011] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

[Spooner *et al.*, 2020] Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific reports*, 10(1):1–10, 2020.

[Wang *et al.*, 2019] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[Zeng *et al.*, 2016] Zhiqiang Zeng, Xiaodong Wang, Jian Zhang, and Qun Wu. Semi-supervised feature selection based on local discriminative information. *Neurocomputing*, 173:102–109, 2016.

[Zhang *et al.*, 2019] Yong Zhang, Qi Wang, Dun-wei Gong, and Xian-fang Song. Nonnegative laplacian embedding guided subspace learning for unsupervised feature selection. *Pattern Recognition*, 93:337–352, 2019.

[Zhao *et al.*, 2008] Jidong Zhao, Ke Lu, and Xiaofei He. Locality sensitive semi-supervised feature selection. *Neurocomputing*, 71(10-12):1842–1849, 2008.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.