Introduction

In supervised machine learning, identifying interactions can be crucial because interactions are scientifically meaningful and interesing in some applications. For instances, @Schwender show that interactions of single nucleotide polymorphisms play an important role in predicting cancer types. @Hamilton2011 suggests there exist interaction effects between political orientation and education levels, holding other background factors constant, in estimating probability of seeing climate change as a threat.

Learning interactions is challenging because including interaction terms might cause overfitting issues and potentially compromise model accuracy performances. Worse, detecting interactions often fall into the high dimensionality curse where there are more features than the observations in the underlying dataset[1]. A common approach to reduce such high dimensionality problem is to apply feature extraction or feature selection methods.

A feature extraction method usually requires transforming feature space to lower dimensionality space, for example, Principal Component Analysis (PCA). However, they are not appropriate for learning interactions since feature transformation can cause loss of information and interpretability. A preferred altenative is feature selection. Feature selection can be loosely defined as determining an optimal set of descriptive features which is associated with the response (target) feature in a dataset by filtering out irrelevant or redundant features [@vural].

There are three main categories of feature selection methods: filter methods, wrapper methods and embedded methods. Filter methods rely on statistical characteristics of data, such as distance and correlation measures, to eliminate poorly associated features. The popular filter methods include Minimum-redundancy-maximum-relevance (mRMR) [@peng] and RELIEF [@Sikonia]. Despite their computational efficiency, filter methods select relevant features by ignoring the performance of the underlying learning algorithm.

Wrapper methods search for the set of descriptive features which optimises a pre-specified accuracy performance measure given a learning model [@kohavi]. Despite a higher computational requirement, wrapper methods often produce more superior performance results. The widely used wrapper methods are Sequential Forward Selection (SFS), Sequantial Backward Selection (SBS), and Genetic Algorithm (GA). A SFS method begins with an empty model (i.e. no descriptive features) and add predictive features sequentially to create an optimal feature set whereas the latter begins with a full set (i.e. with all features) and eliminate one by one to achieve an optimal performance criterion.

In an iterative process, GA generates a population of candidate solutions where each candidate can be mutated or modified to form a new population to be evaluated in the next iteration based on a "fitness" or an objective value function [@GA].The GA is categorised as a meta-heuristic wrapper method. The meta-heuristic approaches often require even higher computational efforts and guarantee no optimal solutions due to their stochastic nature.

Embedded methods incorporate feature selection process when training the underlying learning algorithm. A prominent embedded method is Least Absolute Shrinkage and Selection Operator (LASSO) [@lasso]. LASSO imposes penalty constraints on coefficient estimation of a regression model. It determines the optimal feature set by shrinking the coefficient estimates of irrelevant features to zero. Its variations and extensions include net-elastic models [@glmnet] and group variable selection [@groupLasso].

To apply feature selection methods in detecting interactions, a general approach is to extend the space of feature sets by including interactions themselves. However, direct applications might be prohibitive due to growing computational complexity. To alleviate the computational barrier, there has been a lot of recent development in fitting models with the interactions using embedded methods [@simon; @glinternet; @bien2013; @rajen]. By comparison, there has been little progress for filter and wrapper methods. A recent wrapper-based method is Stepwise cOndtional likelihood variable selection for Discriminant Analysis (SODA)

[1]Given p main effect features or variables, there are $\binom{p}{2}$ *pairwise* interaction terms. By including main effect features, the space of possible feature set grows to $p + \binom{p}{2} = \frac{p(p+1)}{2}$. The space becomes even larger when considering quadratic and higher-order terms.

proposed by @yangli. SODA is built on both forward and backward selections to select predictive main-effect and interactions features to optimise the extended Bayesian Information Criterion (EBIC). We are not aware of any recent development in filter methods for discovering interactions.

The objective of this study is to devise a wrappper method for learning interactions by extending Simultaneous Perturbation Stochastic Approximation in Feature Selection and Ranking (SP-FSR) algorithm introduced by @vural. Given a learning model, this pseudo-gradient descent stochastic algorithm returns a set of predictive features which optimises a specified accuracy performance measure. We propose to utilise the SP-FSR algorithm because @zeren and @vural empirically show that it can produce superior model performances compared to other wrapper methods based on the main-effect features. We wonder if it would yield excellent performances by including interaction features as well. Since we develop our method by applying the SP-FSR algorithm in two major steps, we call it "two-step SP-FSR".

We run computational experiments to compare the performance of our method with other existing approaches. The experiment comprises two types of tasks: regression and binary classification. The target feature is a continuous variable in a regression task whereas it is binary in a classification task. In regression tasks, we compare our method with GA and sequential forward methods with the same learning model - a linear regression. We benchmark our method against a regression-based embedded method as well as a baseline learner, which is a regression without any interaction term. While the wrapper methods can incorporate any learning algorithm, we decide to rely on only one for comparability with the embedded method. Likewise, we set up a classification task in the exact methodology, except that we use a logistic regression as the learning model. In this study, we use the term "interaction" to refer all pairwise interaction features. We shall not consider quadratic and higher-order inteaction terms.

The rest of this thesis paper is organised as follows. The next chapter presents important aspects of "interactions" in statistical context and how these aspects shape the design of our method. In Chapter **??**, we review existing methods for interaction discovery and explain why we select some of them as competitors to two-step SP-FSR. Chapter **??** briefly discusses SP-FSR and delineates how we extend it to identify pairwise interactions. Chapter **??** describes our experiment setup and presents the empirical results. The last chapter concludes with a summary and highlights the direction of our future works.