Kernel-lasso feature expansion method: boosting the prediction ability of machine learning in heart attack

Zongrui Dai1*, Li Jiayi2, Taowen Gong3, and Chenfei Wang4
1,2,4 Westa College, Southwest University, Chongqing, Beibei, 400700, China
3School of Advanced Technology, Xi’an Jiaotong-Liverpool University, Jiangsu, Suzhou, 215000, China
*Corresponding author's e-mail: dzr17723980497@gmail.com

Abstract. Heart attack needs to cause a high degree in modern society, which is now at the top of the list of most common diseases. One-third of all deaths in the world are caused by heart disease, and in our country, hundreds of thousands of people die of heart disease every year. If people can predict heart attack at an early period, they may prevent and cure the potential risk. Therefore, one novel and efficient prediction model is needed. Based on machine learning, we design one novel method in feature selection called kernel-lasso feature expansion to increase the prediction ability. This method considers the potential influence by the curse of dimension and lack of features, which can increase the dimension of the existed data and select the effective features by lasso regression. Compared with other feature selection methods such as Lasso and step regression, the novel method increases the predicted ability of gradient boosting machine, DNN, and SVM models, while SVM with kernel-lasso feature expansion achieves the best performances in Accuracy (0.84(%95CI:0.78-0.89)).

1. Introduction
Heart disease can be one tough task and difficult for people to cure and prevent. Although in most countries, the information and disease features can be stored, scientists lack a powerful tool for data prediction and analysis. While machine learning, which includes many efficient models for prediction, can be a suitable choice. Two main types of machine learning are classification and regression. Classification means prediction for one state or discrete value, while regression aims at prediction for the continuous value. Machine learning has been applied widely applied in medical studies. For instance, machine learning can be used for therapy planning, disease diagnosis, and DNA analysis [1-2], which achieve high accuracy and fast reaction. In our study, we want to know whether the patient can get a heart attack in the future based on the existed data. Therefore, this task can be regarded as a discrete state for 'no heart attack' and 'get heart attack', which means that the classification model can be suitable for the task.

However, existed data from the patents can be lack and general for prediction. Additionally, according to the curse of dimension, too many features can be redundant and inefficient for training. Therefore, a proper number of features can be suitable and efficient for machine learning. Traditional, feature selection such as lasso regression [3], PCA (Principal Component Analysis) [4], and step by step regression [5] can be applied to reduce the dimension of existed data. Although lasso can be applied to cancer selection[6], PCA has been applied to function annotation of bioactive compounds[7], all of them
cannot increase the dimension. When the existed data lack efficient features, feature selection may not increase the ability of machine learning, so that one novel method has to be introduced.

To get better results, kernel-lasso feature expansion is designed to solve the problem. It mainly consists of two parts - feature expansion and feature reduction. In the first step, based on the Gaussian kernel function (formula 1), existed data can be amplified to the high dimensions. Then, lasso regression can be applied to reduce the dimension space of data. Therefore, one novel data with a proper number of features can be obtained. Additionally, this method has a hyperparameter that came from the Gaussian kernel function called 'sigma'.

Based on the data that came from the kernel-lasso feature expansion, machine learning is applied to predict the heart attack. SVM, DNN, and gradient boosting machines have been trained by three cross-validations. In order to compare the kernel-lasso feature expansion with other feature selection methods, each model has been trained with three different data that came from the lasso, step regression, and kernel-lasso feature expansion.

\[
k(||x - xc||) = \exp(-||x - xc||^2 / (2 * \sigma^2))
\]

Formula. 1: The formula for the Gaussian kernel function. \(\sigma\): The width of a function. In this study, \(x\) means one vector of a feature, while \(xc\) means one vector of another features

2. Methodology

2.1. Kernel-lasso feature expansion method
In machine learning tasks, efficient features are required and necessary. However, the ability of machine learning cannot always be improved by adding additional features. Sometimes, redundant features will decrease the generalization ability and produce a poor prediction due to the multicollinearity and noise. Also, the lack of features can cause weak ability in prediction, which can be improved by adding new features. Therefore, the dataset with the proper number of features can boost the performance of the model significantly. Kernel-lasso feature expansion method aims to find the proper number of features from the existed dataset. It combines feature selection and feature boosting together, which can increase the generalization ability of some machine learning models. It can be divided into two parts - feature expansion and lasso reduction. First, according to the existed features (N dimensions), using the RBF kernel function (Formula.1), it can produce some novel features with the number of N!. Second, using lasso regression, which can reduce the dimension through cross-validation, the dimension of the dataset can be decreased to a suitable dimension. Therefore, this method won't be trapped by redundant features and a lack of variables, which can expand the feature space properly.

3. Materials

3.1. Data source
We collected a dataset about some of the characteristics of the individual and their prediction of whether or not they will develop heart disease from Kaggle(https://www.kaggle.com). There are fourteen indicators in the data set that include a prediction of the likelihood of disease. A total of 13 features have an impact on the final result, including age and sex of the patient, exercise-induced angina, ca (number of major vessels), Chest Pain type, resting blood pressure (trtbps), cholestoral in mg/ dl fetched via BMI sensor(chol), fasting blood sugar(fbs), resting electrocardiographic results(rest_ecg) and maximum heart rate achieved(thalach). And there are 303 sets of data for each feature

3.2. Assessment
Confusion matrix consists of TP, FP, FN, and TN, which means True Positive, False Positive, False Negative, and True Negative. Based on the confusion matrix, some assessments can be obtained. For instance, the accuracy, recall, and precision (Formula.2-4).
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)

recall = \frac{TP}{TP + FN} \quad (3)

Precision = \frac{TP}{TP + FP} \quad (4)

Formula 2 - 4: (2): Accuracy is accurate against all data. This is not a good estimate when you have an imbalance of positive and negative samples. (3): Recall refers to the original sample, and it means the probability of being predicted to be a positive sample in a sample that is actually positive. (4): Precision refers to the proportion of correct model prediction among all results in which the model prediction is Positive.

4. Experiment and result

4.1. The stability of hyperparameter in Kernel-lasso feature expansion

Kernel-lasso feature expansion contains one hyperparameter sigma, which determines the width of the kernel function. To validate whether the value of the hyperparameter can influence the performance of model, we apply grid search to test the stability of this method. We test the number of features selected by feature boost method when sigma=0.01, sigma=0.1, sigma=1, sigma=10. The result indicates that the suitable number of features is between 23 and 30(Figure.1). Based on the results, we can point out that the choice of sigma has little effect on the feature selection in Kernel-lasso feature expansion.

Figure 1: The cross-validation of Kernel-lasso feature expansion under different value of sigma. The highest value of AUC indicates the most suitable number of features, which can achieve the highest performance.
4.2. The comparison between different feature selection methods
According to the results, it can be seen that feature boost presents a good performance in both SVM, DNN, and GBM. Its accuracy is higher than those of Lasso and Step Regression in all three models. Under the SVM model, accuracy is higher than those of Lasso and Step Regression in all three models. Its AUC reached 0.89 is slightly lower than the step regression's 0.9. However, the feature boost method outperformed the other two methods in the other two models. For sensitivity, the feature boost method reached 0.89 under SVM, while only 0.78 under the GBM model. The sensitivity of Lasso and Step Regression methods is also at a low level, which means that the sensitivity of the method can be further improved through adjustment. The results showed that the three methods exhibited similar specificity in different models. However, feature boost achieves a higher specificity in deep and GBM methods than the other two methods (Table 1). Therefore, kernel-lasso method may increase the ability of gradient boosting machine (GBM), deep learning (DNN), and Support vector machine (SVM).

Table 1. The performance of each model in cross validation

| Methods+features  | Accuracy      | Sensitivity    | Specificity    | AUC          |
|-------------------|---------------|----------------|----------------|--------------|
| gbm+Feature_boost | 0.81(95CI:0.77-0.86) | 0.78(95CI:0.68-0.88) | 0.85(95CI:0.79-0.92) | 0.85(95CI:0.82-0.88) |
| gbm+Lasso         | 0.80(95CI:0.76-0.86) | 0.78(95CI:0.68-0.88) | 0.84(95CI:0.76-0.93) | 0.85(95CI:0.81-0.88) |
| gbm+step_regression | 0.84(95CI:0.78-0.89) | 0.89(95CI:0.84-0.92) | 0.81(95CI:0.72-0.90) | 0.89(95CI:0.86-0.92) |
| svm+Feature_boost | 0.83(95CI:0.80-0.85) | 0.85(95CI:0.79-0.92) | 0.82(95CI:0.74-0.93) | 0.84(95CI:0.80-0.88) |
| svm+Lasso         | 0.84(95CI:0.80-0.88) | 0.82(95CI:0.75-0.90) | 0.86(95CI:0.76-0.97) | 0.90(95CI:0.87-0.92) |
| svm+step_regression | 0.83(95CI:0.79-0.87) | 0.81(95CI:0.73-0.90) | 0.83(95CI:0.77-0.90) | 0.85(95CI:0.84-0.92) |
| deep+Feature_boost | 0.83(95CI:0.80-0.84) | 0.84(95CI:0.72-0.90) | 0.82(95CI:0.70-0.95) | 0.87(95CI:0.84-0.90) |
| deep+Lasso        | 0.83(95CI:0.78-0.89) | 0.82(95CI:0.77-0.87) | 0.84(95CI:0.79-0.90) | 0.87(95CI:0.83-0.90) |

5. Conclusion
Based on the result, we confirmed that kernel-lasso feature expansion has stable and robust performance in feature selection, whose hyperparameter sigma has low influence. It means that there is no need for scientists to regulate the hyperparameter of sigma in order to gain better efficiency. Moreover, compared this method with other feature selection methods, we figure out that kernel-lasso method may increase the prediction ability of GBM, SVM, and DNN. Although in some cases, this effect may not be significant, it may provide one optional method for people to expand their features, while the data has low quality. In the future, we have to apply this method in some special data sets to test the performance, which have low quality of features. Also, novel combination of feature expansion and feature reduction can be applied.

6. Code release
The R code of kernel-lasso function has already been updated to the Github (https://github.com/Zongrui-Dai/Kernel-lasso-feature-expansion).

References
[1] Ghumbre S.U., Ghatol A.A. (2012) Heart Disease Diagnosis Using Machine Learning Algorithm. In: Satapathy S.C., Avadhani P.S., Abraham A. (eds) Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. Advances in Intelligent and Soft Computing, vol 132. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27443-5_25
[2] J. Jiang; P. Trundle; J. Ren (2010). Medical image analysis with artificial neural networks. , 34(8), 617–631. doi:10.1016/j.compmedimag.2010.07.003
[3] Zou, Hui (2006). The Adaptive Lasso and Its Oracle Properties., 101(476), 1418–1429. doi:10.1198/016214506000000735
[4] Zou, Hui; Hastie, Trevor; Tibshirani, Robert (2006). Sparse Principal Component Analysis., 15(2), 265–286. doi:10.1198/106186006x113430
[5] Liu C., Zhang M., Zheng M., Chen Y. (2003) Step-by-Step Regression: A More Efficient Alternative for Polynomial Multiple Linear Regression in Stream Cube. In: Whang KY., Jeon J., Shim K., Srivastava J. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2003. Lecture Notes in Computer Science, vol 2637. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-36175-8_44

[6] Jing, L., Ng, M. K., & Zeng, T. (2010). Novel Hybrid Method for Gene Selection and Cancer Prediction. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 4(2), 258-265.

[7] Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends in Food Science and Technology,, 83-90.