COMPSCI5100 Machine Learning & Artificial Intelligence for Data Scientists (M)

Case study 2: Feature engineering

Group: 55
Group Member:
Name: Chengkun Yang
Name: Jinke Chen
Name: Yulin Xiao
Name: Lina Chen
Name: Yuge Wang

1. Introduction

In this case study, the study of feature selection was performed. For large and complex datasets, if all features were trained, the time cost of training model can become very long. Moreover, many features that are not highly relevant to the model may also affect the accuracy of the model. Therefore, feature selection is necessary. The utilization of suitable algorithms to select and decrease the features size into datasets with high relevance to the model can minimize the model training time and keep the model accuracy as possible. This time, the feature selection on the extremely high-dimensional brain Electroencephalogram (EEG) data is presented to predict central neuropathic pain (CNP) in patients with spinal cord injury (SCI). In detail, data containing 9 features of 48 electrodes of the 18 participants consist of 8 negatives and 10 positives in classification with 10 repetitions each was collected and analyzed in this case. Leave one subject out cross-validation was used through the SVM and KNN classifiers to determine the accuracy, sensitivity and specificity of the feature selecting methods utilized in this case: the Pearson Correlation Coefficient, the Spearman correlation coefficient, the L1 Regularization Method, and the Recursive Feature Elimination, Cross-Validated method (RFECV). The performances of dimension reduction algorithms were compared through the mean and standard deviation of the parameters listed above at last.

2. Methods

The Pearson Correlation Coefficient and Spearman's Rank Correlation Coefficient were used as screening methods in this study.

The formula of the Pearson Correlation Coefficient is:

$$\rho_{x,y} = (cov(x,y)) \,/\, (\sigma\_x\ \sigma\_y)$$

The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between the two variables.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - y)^2}}$$

The above equation defines the overall correlation coefficient, often represented by the lowercase Greek letter ρ. Estimating the covariance and standard deviation of the sample yields the Pearson correlation coefficient, often represented by the lowercase letter r.

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.

$$\gamma = \rho_{R(x),R(Y)} = \frac{cov(R(X),R(Y))}{\rho_{R(x)}\rho_{R(y)}}$$

The Pearson Correlation Coefficient was then calculated to be 111 and the Spearman's Rank Correlation Coefficient was calculated to be 109.

Use the L1 Regularization Method as the Embedding Method and then use L1 Regularization to arrive at a number of features of 44.

This is followed by a calculation using RFE as the Wrapper method. The Leave-one-out Cross-Validation method was also used. The number of features is 24 after using the RFE method and 18 after using the Leave-one-out Cross-Validation method. The main advantages of KNNs are their theoretical maturity, their simplicity, the fact that they can be used for both classification and regression, and the fact that they can be used for non-linear classification problems. The nonlinear mapping is the theoretical basis of the SVM method, which uses an inner product kernel function instead of a nonlinear mapping to a higher dimensional space. The optimal hyperplane for the partitioning of the feature space is the goal of the SVM, and the idea of maximizing the classification margins is the core of the SVM method. Two Classifiers, KNN and SVM are chosen. The results are then compared.

3. Results

In this part, four feature selection methods: Pearson correlation coefficient, Spearman correlation coefficient, RFECV and L1 regularization were tested, the SVM and KNN classifiers were utilized through the Leave one subject out cross-validation to determine and compare the performance of accuracy, sensitivity and specificity for the four methods.

Firstly, the definition of precision, sensitivity and specificity can be expressed as follow:

Table 1 Performance metrics

Reality	True	False
Positive	TP	TN
Negative	TF	FN

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

For these three performance evaluation indicators, the higher the values are, the better the performance can be indicated.

The features were selected using the filtering method, wrapping method and embedding method. In this case, the filtering method tested included the method of Pearson correlation coefficient and Spearman correlation coefficient.

1) For the Pearson correlation coefficient, the number of features was reduced to 111.
2) For the Spearman correlation coefficient, the number of features was reduced to 109.
3) For the L1 regularization, the number of features was reduced to 44.
4) For the RFECV, the number of features was reduced to 24.

The means and standard deviations of the two classifiers obtained using these four methods are shown in the Figure.

Means

Out[26]:

Metrics	knn	svm
accuracy	0.877778	0.905556
sensitivity	0.800000	0.837500
specificity	0.940000	0.960000

Figure 1 Means of Pearson correlation coefficient

Means

Out[33]:

Metrics	knn	svm
accuracy	0.855556	0.638889
sensitivity	0.850000	0.612500
specificity	0.860000	0.660000

Figure 2 Means of Spearman correlation

Standard deviations

Out[48]:

Metrics	knn	svm
accuracy	0.237788	0.132521
sensitivity	0.161536	0.121835
specificity	0.094340	0.140000

Figure 5 Standard deviations of Pearson

Standard deviations

Out[34]:

Metrics	knn	svm
accuracy	0.265041	0.292762
sensitivity	0.264575	0.169097
specificity	0.265330	0.361109

Figure 7 Standard deviations of L1 regularization

coefficient

Means

Out[40]:

Metrics	knn	svm
accuracy	0.666667	0.883333
sensitivity	0.437500	0.875000
specificity	0.850000	0.890000

Figure 3 Means of L1 regularization

Means

Out[47]:

Metrics	knn	svm
accuracy	0.711111	0.872222
sensitivity	0.487500	0.862500
specificity	0.890000	0.880000

Figure 4 Means of RFECV

Standard deviations

Out[27]:

Metrics	knn	svm
accuracy	0.161780	0.135287
sensitivity	0.173205	0.165359
specificity	0.120000	0.066332

Figure 6 Standard deviations of Spearman

Standard deviations

Out[41]:

Metrics	knn	svm
accuracy	0.230940	0.083333
sensitivity	0.099216	0.082916
specificity	0.111803	0.083066

Figure 8 Standard deviations of RFECV

For the KNN classifier, the mean value of the performance index of accuracy is the best using the Pearson correlation coefficient method, followed by the wrapping method. The result obtained by using the Spearman correlation coefficient method is the lowest, indicating that the Pearson correlation coefficient method is used. The classifier has a

higher probability of pairing samples; sensitivity indicates the proportion of all positive cases that are paired, and measures the ability of the KNN classifier to identify positive cases. The mean value of sensitivity obtained by using the wrapping method is the highest, indicating that Using the package method KNN has the best ability to identify positive examples; for specificity, it represents the proportion of all negative examples that are paired, which measures the ability of the classifier to identify negative examples, and is obtained using the Pearson correlation coefficient The mean value of is the highest, indicating that the method has a stronger ability to identify negative examples.

For the SVM classifier, the mean value of precision and specificity is higher using the Pearson correlation coefficient method, indicating that using this method, the SVM classifier has a higher probability of pairing samples and a stronger ability to identify negative examples; Using the Spearman correlation coefficient, the SVM classifier has a stronger ability to identify positive examples.

For the KNN classifier, when evaluating the performance of the accuracy, the standard deviation of the method using the Pearson correlation coefficient is the smallest, indicating that the stability of the method is the best; when evaluating the sensitivity, the standard of the Spearman correlation coefficient is used When the difference is the smallest, the Spearman method gives the best stability; when specificity is assessed, the embedding method gives the best stability.

For the SVM classifier, the performance evaluation of accuracy, sensitivity and specificity is performed. Using the Spearman correlation coefficient method, the variance of these three performances is the smallest, indicating that the Spearman method has the best stability for these three performances.

4. Discussion

From the results, the Pearson correlation coefficient method in the filtering method outperforms the other feature selection methods in this experiment because the Pearson correlation coefficient method performs better overall on the SVM classifier and KNN classifier, followed by the Spearman correlation coefficient method.

Since the filtering method is not based on the model but on the general performance of the features when selecting features, the filtering method is more general, can effectively avoid overfitting, and is very suitable as a pre-screener for features.

At the same time, because the feature selection of the filtering method is completely

independent of the specific learning algorithm, the selected feature subset is usually lower than the wrapper and embedding methods in terms of classification accuracy.

However, based on this experiment, the accuracy of the filtering method used is higher than that of the wrapper and embedding methods, so it is considered that features selected by L1 regularization and RFECV may be not enough, and the number of features removed is too large, resulting in a consequence value that is more biased towards local performance rather than overall performance. The filtering method selects more features and thus performs better, but it is also possible that the selected features are similar or irrelevant to the target, resulting in redundancy.

Also, comparing the KNN and SVM classifiers, it can be known that the SVM classification is better: the means are higher, and the standard deviations are lower overall. Since KNN does not have a training process, each sample needs to be considered during the training process, and the prediction efficiency is low when the training and test sets are large. SVM classifies the test set directly by training the model, and is suitable for dealing with small samples, nonlinear, and high-dimensional pattern recognition.