Title: Case study 1 Model selection for clustering

Name: Akash Ananda Kumar Murali **ID: 2768146M**

Name: Hao Chen **ID: 2691051C**

Name: Jiachen Dong **ID: 2789896D**

Name: Wenbin Shang **ID: 2743882S**

Name: Xinyu Guan **ID: 2764666G**

1.Introduction

1.1 background and data

In this case we focus on preprocessed brain EEG data sets. The data set here is from SCI patients recorded while resting with eyes closed (EC) and eyes opened (EO). So in this case, the data set mentions about 48 electrodes recording electrical activity of the brain 250HZ, and which has two possibility classes where subjects will/will not develop neuropathic pain within 6 months. So there were totally 18 subjects, in which 10 developed pain and 8 didn't develop pain. The data is already undergone some preprocessing like signal denoising and normalization, temporal segmentation, frequency band power estimation, normalization with respect to total band power and features include normalized alpha, theta band power while eyes are closed, eyes opened, and taking the ratio of eo/ec.

1.2 statement of task

Our main task in this case is to choose four different feature selection methods(two from filtering method, wrapper method, and embedding method) and use two different classifiers (SVM and KNN) to deal with the data set. Then perform Leave-one-group-out cross-validation, use cross-validation to optimize hyper-parameter values of the dataset and use the classifiers to train the data set without feature selection.

2. Methods

This section introduces the three feature selection methods and two classifiers, describes the content and steps of each method, and describes the role of important parameters in the code.

2.1 Filtering method

Filter methods are used as a preprocessing step. The feature selection is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here. The filter method ranks each feature by some uni-variate metric and then selects the highest-ranking features.[1]

There are some advantages to use filtering method:

a. Filter methods are model agnostic

b. Depends completely on features in the data set

c. Computationally faster

d. Based on different statistical methods

There are also some disadvantages, The filter method looks at individual features for identifying its relative importance. A feature may not be useful on its own but maybe an important influencer when combined with other features. Filter methods may lose such features.

2.1.1 VarianceThreshold

Filtering is often used in the pre-processing of data for machine learning, and variance filtering is one of the filtering methods. Variance filtering is the process of filtering out features that have a small variance. For example, if the variance of a feature is very small, it means that the sample does not differ in this feature, and most of the values in the feature are the same, or even the whole feature has the same value. So a filtering threshold can be

set to filter out those features with small variance, thus achieving the purpose of feature filtering.

Parameter: The threshold value passed in is the most critical parameter, which determines the strength of your filtering of the data. The more stringent the filtering the less data you get; the simpler the filtering the more data you get. If it's too simple, it's the original data that hasn't been moved. Here we are using a median as a parameter, effectively preventing potential pitfalls such as overfitting.

2.1.2 Chi-Square Test

The chi-square filter is a correlation filter specifically for discrete labels (i.e. classification problems)The chi-square test sklearn.feature_selection.chi2 calculates the chi-square statistic for each non-negative feature and label. The features are then ranked according to their statistics from highest to lowest.In combination with the sklearn.feature_selection-SelectKBest class, the K most relevant features are selected
Note: If the chi-square test detects that most of the values in a feature are the same, we are prompted to filter for variance using the variance mentioned earlier
parameter： The chi-square test chi2 returns two statistics, the chi-square value and the p-value, where the chi-square value is difficult to define the range and the p-value, generally using 0.05 or 0.01 as the level of significance, i.e. the boundary for the p-value judgement. Finally we use chivalue_value.shape[0]-(pvalues_chi_value>0.05).sum() to calculate our most critical and most appropriate k-value, using the k-value to pass into SelectKBest, and the filtering is complete.

2.2 Wrapper method

In this method, a subset of features are selected and train a model using them. Based on the inference that we draw from the previous model, we decide to add or remove features from the subset.[2]

There are some advantages to use wrapper method:

a. Less prone to local optima

b. Interacts with the classifier

c. Models feature dependencies

d. Higher performance accuracy than filter

There are also some disadvantages, such as computationally intensive, discriminative power, lower shorter training times, and classifier dependent selection.

2.3 Embedding method

In this method, feature selection is integrated or built into the classifier algorithm. During the training step model, the classifier adjusts its internal parameters and determines the appropriate weights/importance given for each feature to produce the best classification accuracy.[3]

There are some advantages to use embedding method:

a. Interacts with the classifier

b. The models feature dependencies better computational complexity than the wrapper

c. Higher performance, accuracy than filter

d. Preserving data characteristics for interpretability

There are also some disadvantages, such as Classifier dependent selection, and Consider the dependence among features.

2.4 Support Vector Machine(SVM)

SVM can be used in both classification and regression problems. But it is mainly used for classification. In the SVM algorithm, first it plots each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, it performs classification by finding the hyper-plane that differentiates the two classes very well.[4]

There are some advantages to use SVM:

a. more effective in high dimensional spaces.

b. effective in cases where the number of dimensions is greater than the number of samples.

c. relatively memory efficient

There are also some disadvantages, such as SVM is not suitable for large data sets, and does not perform very well when the data set has more noise.

2.5 K-Nearest Neighbour(KNN)

The KNN can be used for classification and regression problems. The KNN algorithm is a non-parametric, supervised learning classifier. it uses proximity to make classifications or predictions about the grouping of an individual data point.[5]

There are some advantages to use KNN:

a. No Training Period

b. easy to implement

c. new data can be added seamlessly

There are also some disadvantages, such as it does not work well with large dataset, Need feature scaling, and Sensitivity to noisy data, missing values and outliers.

3. Results

3.1 Feature Filtering method & Classifier Method

The data shown below is filtered by each of the four feature selection methods and then trained on the filtered data by two classification methods. Two different models, KNN and SVM, were obtained.

After the two models were derived we then scored a total of 8 different cases for 4 * 2

More specifically, each of the following filtering methods corresponds to two different classification algorithms, where the KNN algorithm has 4 different K values and the SVM algorithm has 3 different kernels. In total, there are 7 scoring functions and the optimal hyperparameters are selected separately.

Filtering method：VarianceThreshold

when the n_neighbors is 1 Knn Accuracy: 0.6111111111111112

when the n_neighbors is 2 Knn Accuracy: 0.6666666666666666

when the n_neighbors is 3 Knn Accuracy: 0.6944444444444444

when the n_neighbors is 4 Knn Accuracy: 0.6388888888888888

the final knn n_neighbors is 3

The score of linear kernel is : 0.722222

The score of rbf kernel is : 0.694444

The score of poly is kernel : 0.611111

final svm kernel hyper-parameter value is: linear

Filtering method：Chi-Square Test

when the n_neighbors is 1 Knn Accuracy: 0.7222222222222222

when the n_neighbors is 2 Knn Accuracy: 0.6944444444444444

when the n_neighbors is 3 Knn Accuracy: 0.7222222222222222

when the n_neighbors is 4 Knn Accuracy: 0.7222222222222222

the final knn n_neighbors is 1

The score of linear kernel is : 0.750000

The score of rbf kernel is : 0.694444

The score of poly is kernel : 0.638889

final svm kernel hyper-parameter value is: linear

Wrapper method

when the n_neighbors is 1 Knn Accuracy: 0.7222222222222222

when the n_neighbors is 2 Knn Accuracy: 0.7222222222222222

when the n_neighbors is 3 Knn Accuracy: 0.75

when the n_neighbors is 4 Knn Accuracy: 0.7222222222222222

the final knn n_neighbors is 3

The score of linear kernel is : 0.555556

The score of rbf kernel is : 0.722222

The score of poly is kernel : 0.694444

final svm kernel hyper-parameter value is: rbf

Embedding method

when the n_neighbors is 1 Knn Accuracy: 0.6111111111111112

when the n_neighbors is 2 Knn Accuracy: 0.8055555555555556

when the n_neighbors is 3 Knn Accuracy: 0.5833333333333334

when the n_neighbors is 4 Knn Accuracy: 0.6666666666666666

the final knn n_neighbors is 2

The score of linear kernel is : 0.611111

The score of rbf kernel is : 0.638889

The score of poly is kernel : 0.583333

final svm kernel hyper-parameter value is: rbf

3.2 Leave-one-group-out cross-validation and Score

The output below shows the data selected by the feature selection method using the Leave-one-group-out cross-validation method to classify a total of 18 groups of 10 elements each, followed by a score judgement.

A summary of the scores from both the KNN and SVM functions is shown below, and the average was selected as the final evaluation.

Scoring all the combos

knn final score: [0.5 0.6 0.7 0.6 0.8 0.6 0.8 0.5 0.8 0.7 0.9 0.5 0.7 0.7 0.6 0.6 0.6 0.6]

The mean value: 0.6555555555555554

svm final score: [0.9 0.6 0.9 0.6 0.6 0.7 0.9 1. 0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.7 1.]

The mean value: 0.8

knn final score: [0.8 0.6 0.8 0.5 0.9 0.7 0.6 0.9 0.9 0.6 0.5 0.7 0.6 0.6 0.7 0.9 0.6 0.9]

The mean value: 0.7111111111111111

svm final score: [0.7 1. 0.7 0.6 0.9 0.8 0.8 0.8 0.8 0.8 0.7 0.8 0.7 0.5 0.8 0.9 0.8 1.]

The mean value: 0.7833333333333334

knn final score: [0.8 0.8 0.9 0.6 0.7 0.8 0.9 0.7 0.9 0.9 1. 0.7 0.7 0.6 0.6 0.9 0.8 0.7]

The mean value: 0.7777777777777778

svm final score: [0.7 0.7 0.6 0.6 0.8 0.6 0.8 0.7 0.7 0.7 0.6 0.4 0.6 0.5 0.5 0.7 0.5 0.7]

The mean value: 0.6333333333333333

knn final score: [0.6 0.5 0.7 0.7 0.8 0.6 0.7 0.8 0.8 0.8 0.8 0.3 0.8 0.7 0.5 0.7 0.6 0.5]

The mean value: 0.6611111111111111

svm final score: [0.6 0.6 0.7 0.6 0.7 0.6 0.6 0.6 0.7 0.6 0.5 0.6 0.5 0.5 0.5 0.5 0.5 0.5]

The mean value: 0.5777777777777778

3.3 Training on the original data and comparing scores

We also used the two classifiers to evaluate the data set without the feature selections. The result is shown below.

The score of SVM is : 0.861111

KNN Accuracy: 0.6944444444444444

4. Conclusion

For variance selection and chi-square selection:
The optimal SVM kernel in variance selection is linear, and the optimal k-value of knn is 3.

Compared with the original data, the accuracy of svm decreased slightly after variance selection, from 0.86 to 0.72, and the accuracy of knn remained unchanged from 0.69 to 0.69

The optimal SVM kernel in chi-square selection is linear, and the optimal knn k value is 1
Compared with the original data, the accuracy of svm decreased slightly after chi-square selection, from 0.86 to 0.75, and the accuracy of knn increased slightly from 0.69 to 0.72

For embedding method and wrapper method:
The KNN classifier scoring for the two filtering method is a little bit lower than the wrapper method and the embedding method. The SVM classifier scoring for the embedding method is lower than the other three methods. Each classifier scoring using in wrapper method is much higher than other methods, due to this, using wrapper method in this data set is much better than any other feature selection methods.

For the data after Leave-one-group-out cross-validation processing:
After the four feature selection methods pass the Leave-one-group-out cross-validation, the highest score is the variance selection method and uses svm, and the score of 0.8 is the best

Reference:

[1]
https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

[2]
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

[3]
https://machinelearningmastery.com/calculate-feature-importance-with-python/

[4]
https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[5]
https://www.ibm.com/topics/knn