**Feature Extraction, Feature Selection and Dimension reduction**

**Introduction**

In machine learning, the outcome of regression or classification models depends on features, continuous or categorical variables. In real world scenarios, a data set may have thousands of features, and all features are not equally important for a definite outcome. It is not always practical to work with all possible features due to expense. Feature extraction and selection are used to identify and remove irrelevant and redundant features. The goal of feature selection is to remove the redundant feature(s) and select relevant feature for further analysis, for example, classification[1]. Dimension reduction is possible as some features don't have a significant impact on the target variable, and some features are strongly correlated with each other[2].

The aims of this study are given below.

* Facilitating data visualization and data understanding,
* Improving Efficiency
* Reducing the complexity of a model
* Reducing the measurement and storage requirements,
* Improving Prediction Performance of the predictors in the model
* Defying the curse of dimensionality to improve prediction performance

There are three types of feature selection method. These are-

1. Filter methods
2. Wrapper method
3. Embedded method

Here, the filter method of feature selection is used, thus, in the following sections, only the filter method will be discussed.

**Filter Methods**

Filter methods use attribute evaluator and a ranker to rank all the features in your dataset without considering learning algorithms that use for decision making. The following figure, Fig 1, demonstrates filter method for feature selection.

Features

Filter Method

Learning Algorithms

Fig 1: Filter Method

Many methods use only for either categorical or continuous features and some for both types of features. All features of our dataset are continuous values, and the target variable is categorical. We will discuss the dataset in the following section. Thus, Here, filter methods that can analyze continuous data are used for feature selection and dimension reduction. The list of filter methods applied for the dataset are given below.

1. Point Biserial Correlation
2. Relief algorithm
3. Fisher Score
4. Pearson correlation
5. Linear Discriminant Analysis (LDA)

**Point Biserial Correlation Coefficient**

All methods were other than Point-biserial correlation discussed in the class lectures. Pearson correlation is applicable to continuous variables. In our case, the target variable is categorical variable with two class and features are continuous variables. In this case, Point-biserial correlation coefficient is used to determine the correlation between variables. In this project, Point-biserial correlation coefficient is used to evaluate the correlation between the target variable and features. The following equation is used to determine Point-biserial correlation coefficient.

Where, and are the mean values of continuous variables in group 1 and group 2. is the number of instances. are number of instance in group 1 and group 2, respectively. is the instance of continuous variable and is the mean of continuous variable. The higher value of variable indicates the higher correlation.

**Description of Dataset**

The data set of breast cancer collected from diagnosis of the University of Wisconsin Hospitals. The diagnosis procedure begins by obtaining a small drop of fluid from a breast tumor using a fine needle. The images of the drop of fluid were analyzed. In this way, a total of 569 samples were collected and analyzed. From a set of image parameters, 10 parameters were selected which best separate benign from malignant samples. Ten features are computed for each sample: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension [3, 4]. The mean value, standard error (SE), the extreme value of each parameter are features of the dataset. So, the dataset has 30 features, and the classes of the target variable (benign and malignant). The data set is adapted from UCI Machine Learning Repository data.

**Results**

Firstly, by point biserial correlation coefficient method, all features are selected which have a correlation with the target variable above 30%. We get 23 features. Secondly, We selected 23 features by using the Fisher score and Relief algorithm. By analyzing the outcomes of three methods which select features based on the relationship with the target variable it is found that among the selected features 21 features are common. In the following step, to reduce features according to correlation among features, the selected 21 features are investigated by applying the Pearson correlation. We considered two features are redundant features that have a correlation equal to or above 0.8. From two redundant features, we selected the feature that has a comparatively high correlation with target variables. Finally, wee selected from 21 features10 features that have a significant correlation with the target variable and correlation with other features below 0.8.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
| SL. No. | Point Biserial Correlation | Fisher Score | Relief Algorithm |
| 28 | concave points\_worst | concave points\_worst | concave points\_worst |
| 23 | perimeter\_worst | perimeter\_worst | compactness\_worst |
| 8 | concave points\_mean | concave points\_mean | smoothness\_worst |
| 21 | radius\_worst | radius\_worst | concavity\_worst |
| 3 | perimeter\_mean | perimeter\_mean | compactness\_mean |
| 24 | area\_worst | area\_worst | compactness\_mean |
| 1 | radius\_mean | radius\_mean | concavity\_mean |
| 4 | area\_mean | area\_mean | concave points\_mean |
| 7 | concavity\_mean | concavity\_mean | symmetry\_worst |
| 27 | concavity\_worst | concavity\_worst | fractal\_dimension\_worst |
| 6 | compactness\_mean | compactness\_mean | texture\_mean |
| 26 | compactness\_mean | compactness\_mean | smoothness\_mean |
| 11 | radius\_se | radius\_se | fractal\_dimension\_mean |
| 13 | perimeter\_se | perimeter\_se | symmetry\_mean |
| 14 | area\_se | area\_se | perimeter\_worst |
| 22 | texture\_worst | texture\_worst | radius\_worst |
| 25 | smoothness\_worst | smoothness\_worst | area\_worst |
| 29 | symmetry\_worst | symmetry\_worst | perimeter\_mean |
| 2 | texture\_mean | texture\_mean | radius\_mean |
| 18 | concave points\_se | concave points\_se | area\_mean |
| 5 | concave points\_se | concave points\_se | compactness\_se |
| 9 | symmetry\_mean | symmetry\_mean | perimeter\_se |
| 30 | fractal\_dimension\_worst | fractal\_dimension\_worst | concave points\_se |

|  |  |
| --- | --- |
| SL. No. | Features that are common in three methds |
| 28 | concave points\_worst |
| 23 | perimeter\_worst |
| 8 | concave points\_mean |
| 21 | radius\_worst |
| 3 | perimeter\_mean |
| 24 | area\_worst |
| 1 | radius\_mean |
| 4 | area\_mean |
| 7 | concavity\_mean |
| 27 | concavity\_worst |
| 6 | compactness\_mean |
| 26 | compactness\_mean |
| 13 | perimeter\_se |
| 22 | texture\_worst |
| 25 | smoothness\_worst |
| 29 | symmetry\_worst |
| 2 | texture\_mean |
| 18 | concave points\_se |
| 5 | concave points\_se |
| 9 | symmetry\_mean |
| 30 | fractal\_dimension\_worst |

|  |  |
| --- | --- |
| SL. No. | Features |
| 28 | concave points\_worst |
| 21 | radius\_worst |
| 1 | radius\_mean |
| 13 | perimeter\_se |
| 22 | texture\_worst |
| 25 | smoothness\_worst |
| 29 | symmetry\_worst |
| 18 | concave points\_se |
| 9 | symmetry\_mean |
| 30 | fractal\_dimension\_worst |

**Conclusions**

In this project, we compare three (03) feature selection methods for breast cancer dataset. From the outcomes of three methods. We recommend using those common features for learning algorithms as each feature selection method has its own shortcomings. Finally, PCA is used to reduce dimension fo dataset to three for visualization.

1. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.

2. Carreira-Perpinán, M.A., *A review of dimension reduction techniques.* Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09, 1997. **9**: p. 1-69.

3. Mangasarian, O.L., W.N. Street, and W.H. Wolberg, *Breast cancer diagnosis and prognosis via linear programming.* Operations Research, 1995. **43**(4): p. 570-577.

4. Street, W.N., W.H. Wolberg, and O.L. Mangasarian. *Nuclear feature extraction for breast tumor diagnosis*. in *Biomedical image processing and biomedical visualization*. 1993. International Society for Optics and Photonics.