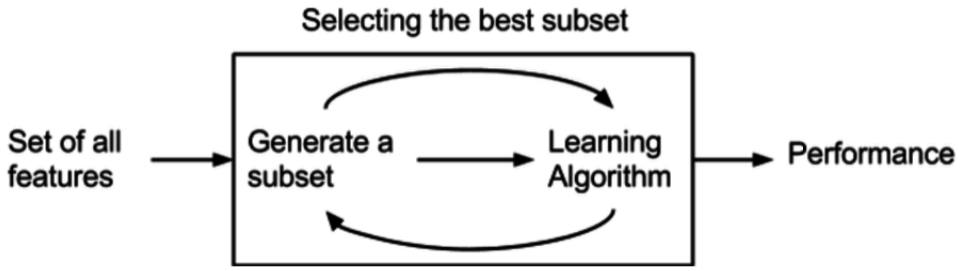
Feature Selection and Extraction Analysis Implementation Issues

Feature selection or variable selection is the process of selecting a subset of relevant features for use in model construction. Consider a dataset of students in college over a period of few years. Let us say we have 2 million rows with 100 features in the dataset. Feature selection helps us pick the relevant features to create a model with instead of using all the 100 features. Assume that two of the features, name and ID, used to predict the marks allotted to a student in an exam. Since we require only one of these features for identification we get rid of the name attribute. If you remove this attribute, the speed and accuracy of the model will increase. The above example is one of the many situations where feature selection helps get rid of irrelevant features.

The following are the implementation issues associated with feature selection:

- Filter methods is a class of feature selection that selects variables regardless of their model based on general features like the correlation with variable to predict. The issue with this method is that it tends to select redundant variables because it does not consider relationships between variables.
- Wrapper methods is a class of feature selection that evaluate subsets of variables which allows the detection of possible interactions between variables. The flaw with this method is the risk of increased overfitting when the number of observations is insufficient.
- Another flaw associated with the wrapper method is the

significantly large computation time when the number of variables is large.



The figure above depicts how the Wrapper method works for feature selection

- Variance threshold is a baseline approach of feature selection. It removes all features whose variance does not meet some threshold. For instance, if a public health dataset consists of 96% of its observations being male and 35 years old, then age and gender can be eliminated without major loss of information. If your problem requires dimensionality reduction, applying variance thresholds is not sufficient. You will have to manually set a variance threshold, which is challenging.
- Correlation threshold removes features whose values change very similarly to another value, since these features provide redundant information. For instance, if you had a public health dataset which had its values in the metric and the SI system, you can safely remove one of these. Just like the method above, you must manually set a correlation threshold, which is challenging. Therefore, algorithms with built in thresholds are preferred over this method.
- Genetic algorithms (GAs) are a method for function optimization based on the mechanics of natural genetics and biological evolution. For instance, consider that 'genes' represent individual

features and the 'organism' represents a candidate set of features. Each organism in the 'population' is graded on a fitness score such as model performance on a hold-out set. The fittest organisms survive and reproduce, until the population converges on a solution. GAs increase the complexity of your implementation. The mutation operator likes adding a randomly generated number to a parameter of an individual of the population. Therefore, the GA has slow convergence.

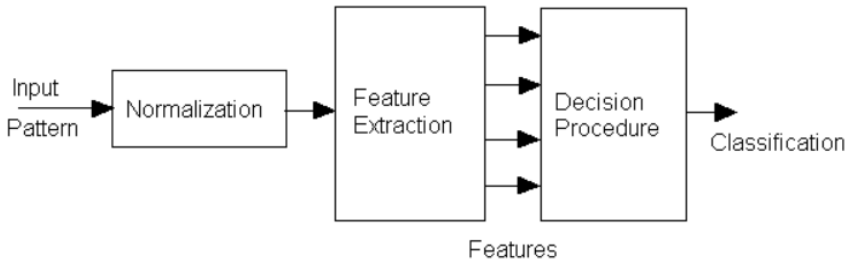
- Stepwise search is a supervised feature selection method based on sequential search. It works forwards and backwards. The most fatal flaw with this method is that it is a greedy algorithm and cannot account for future effects of each change.

Feature extraction serves the purpose of not only reducing dimensionality, but also of extracting useful information hidden in signals while avoiding redundant information. For example, consider an image. We know an image consists of pixels and each pixel can have an 8-bit value. So, even a 640 x 480 image will have 640 x 480 x 8 bits of information. This is too much data for a computer to make sense of directly. We can use feature extraction to figure out which parts of an image are distinctive, such as lines, corners, and special patches that can describe an image uniquely.

The following are the implementation issues associated with feature selection:

- Feature extraction in sentiment analysis faces distinct issues such as large feature space problems, redundancy, domain dependency, difficulty in implicit feature identification, and limited work on Lexico-structural features (Asghar, Khan, Ahmad, Kundi, 2014).

- High-dimensionality/ large feature sets cause performance degradation due to computational problems and thus need feature selection is then required.
- Performance of clustering-based feature extraction techniques is domain dependent, and hence results in creating cross domain and generalization problems.
- Ability of hybrid methods to overcome problems arising from redundancy needs further research since it is not yet confirmed.



The figure above depicts the feature extraction process

- After feature extraction, you may end up with features which may not be different from each other and hence end up neglected. But, discarding a feature which has definite importance in classifying that object may affect the predictive accuracy of the classifier.
- Autoencoders are neural networks that are trained to reconstruct their original inputs. A drawback is that since they are neural networks they require more data to train. They are not used as general-purpose dimensionality reduction algorithms (Nixon, Aguado, 2002).
- Linear discriminant analysis (LDA) creates linear combinations of your original features but does not maximize explained

variance. It is a type of feature extraction, which is not easy to interpret, and you are required to manually set the number of components to retain.

- Another limitation of the LDA method is that it requires labeled data, which makes it more situational.
- Principal Component Analysis (PCA) is an unsupervised algorithm that creates linear combinations of original features. This is a type of feature extraction method whose new principal components are not interpretable.
- Another issue with PCA as a technique under feature extraction is that you will have to manually set a threshold for cumulative explained variance.

Asghar, M. Khan, A. Ahmad, S. Kundi, F. A Review of Feature Extraction in Sentiment Analysis. 2014. Retrieved from: http://www.statmt.org

Nixon, M. Aguado, A. Feature Extraction and Image Processing. 2002. Retrieved from: https://www.ppgia.pucpr.br