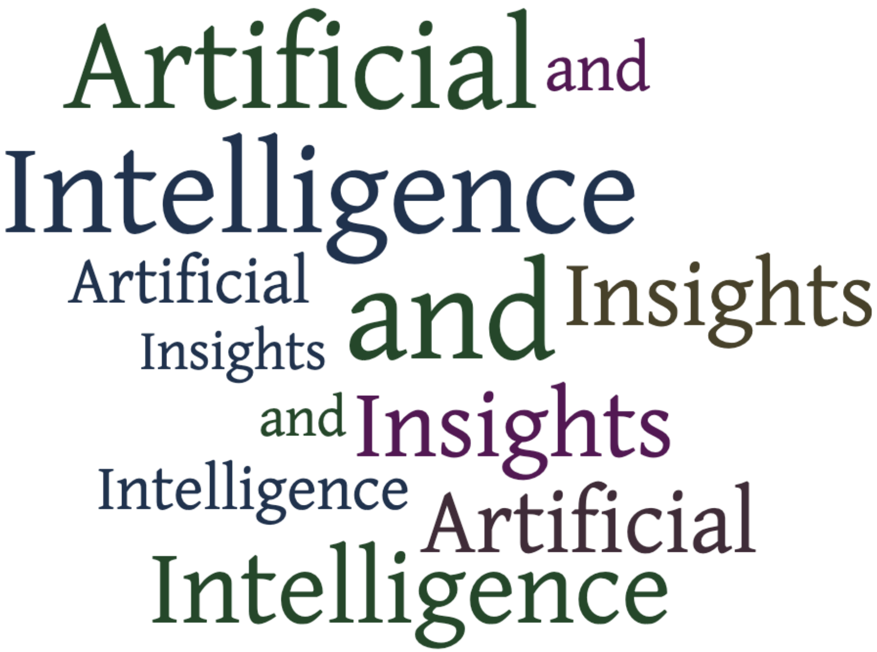
SWINBURNE UNIVERSITY OF TECHNOLOGY



INF80051 AI and insights

Module 3 lecture

Feature Selection

What is feature selection

In previous lectures, we considered all features are available prior to the design of the classifier. However, a major problem in developing AI algorithms is the so-called curse of dimensionality.

The number of features at the disposal of the designer of a classification system is usually very large. This number can easily reach the order of a few dozens or even hundreds.

What is feature selection

There is more than one reason to reduce the number of features to a sufficient minimum. Computational complexity is the obvious one.

A related reason is that, although two features may carry good classification information when treated separately, there is little gain if they are combined into a feature vector because of a high mutual correlation.

What is feature selection

Thus, complexity increases without much gain. Another major reason is that imposed by the required generalization properties of the classifier.

The higher the ratio of the number of training patterns N to the number of free classifier parameters, the better the generalization properties of the resulting classifier.

What is feature selection

A large number of features are directly translated into a large number of classifier parameters. Thus, for a finite and usually limited number N of training patterns, keeping the number of features as small as possible is in line with our desire to design classifiers with good generalization capabilities.

What is feature selection

Furthermore, the ratio N/l enters the scene from another nearby corner. One important step in the design of a classification system is the performance evaluation stage, in which the classification error probability of the designed classifier is estimated. We not only need to design a classification system, but we must also assess its performance.

What is feature selection

In summary the task of feature selection can be translated into:

Given a number of features, how can one select the most important of them so as to reduce their number and at the same time retain as much as possible of their class discriminatory information? The procedure is known as feature selection or reduction.

What is feature selection

If we selected features with little discrimination power, the subsequent design of a classifier would lead to poor performance. If information-rich features are selected, the design of the classifier can be greatly simplified.

In a more quantitative description, we should aim to select features leading to large between-class distance and small within-class variance in the feature vector space.

What is feature selection

This means that features should take distant values in the different classes and closely located values in the same class. To this end, different scenarios will be adopted. One is to examine the features individually and discard those with little discriminatory capability.

A better alternative is to examine them in combinations. Sometimes the application of a linear or nonlinear transformation to a feature vector may lead to a new one with better discriminatory properties.

Feature selection based on statistical hypothesis testing

A first step in feature selection is to look at each of the generated features independently and test their discriminatory capability for the problem at hand.

Although looking at the features independently is far from optimal, this procedure helps us to discard easily recognizable “bad” choices and keeps the more elaborate techniques, which we will consider next, from unnecessary computational burden.

Feature selection based on statistical hypothesis testing

Let x be the random variable representing a specific feature. We will try to investigate whether the values it takes for the different classes, say 1, 2, differ significantly.

To give an answer to this question, we will formulate the problem in the context of statistical hypothesis testing. That is, we will try to answer which of the following hypotheses is correct:

- H1: The values of the feature differ significantly
- H0: The values of the feature do not differ significantly

Feature selection based on statistical hypothesis testing

Let *x* be the random variable representing a specific feature. We will try to investigate whether the values it takes for the different classes, say 1, 2, differ significantly. To give an answer to this question, we will formulate the problem in the context of statistical hypothesis testing.

That is, we will try to answer which of the following hypotheses is correct:

- H1: The values of the feature differ significantly
- H0: The values of the feature do not differ significantly

H0 is known as the null hypothesis and H1 as the alternative hypothesis. The decision is reached on the basis of experimental evidence supporting the rejection or not of H0.

OTHER TECHNIQUES

- The Receiver Operating Characteristics (ROC) curve
- Class separability measures

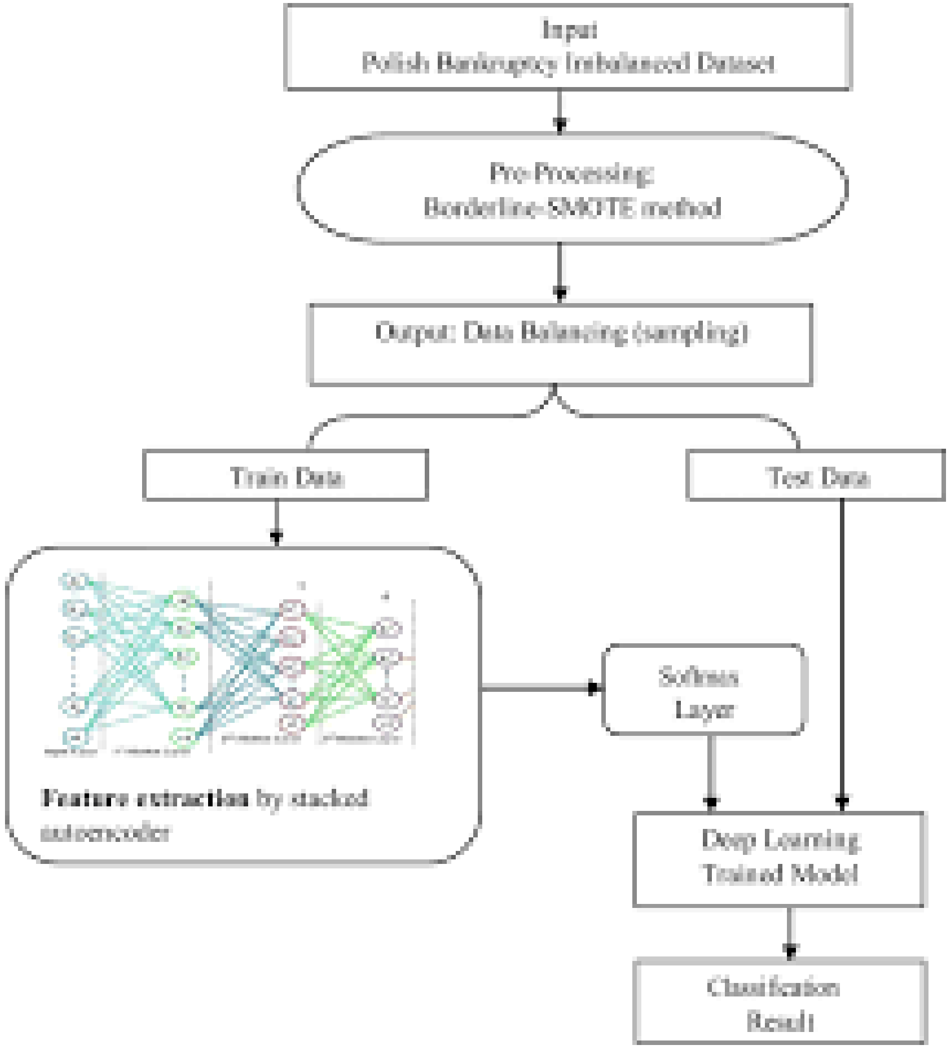
Bankruptcy prediction

Bankruptcy prediction is a binary classification problem in financial distress prediction, which aims at assigning new observations to two pre-defined decision classes. For example, bankruptcy prediction models are used to predict the likelihood that the loan customers will go bankrupt whereas. In credit scoring models are used to determine whether the loan applicants should be classified into a high risk or low risk group.

Feature selection for bankruptcy prediction

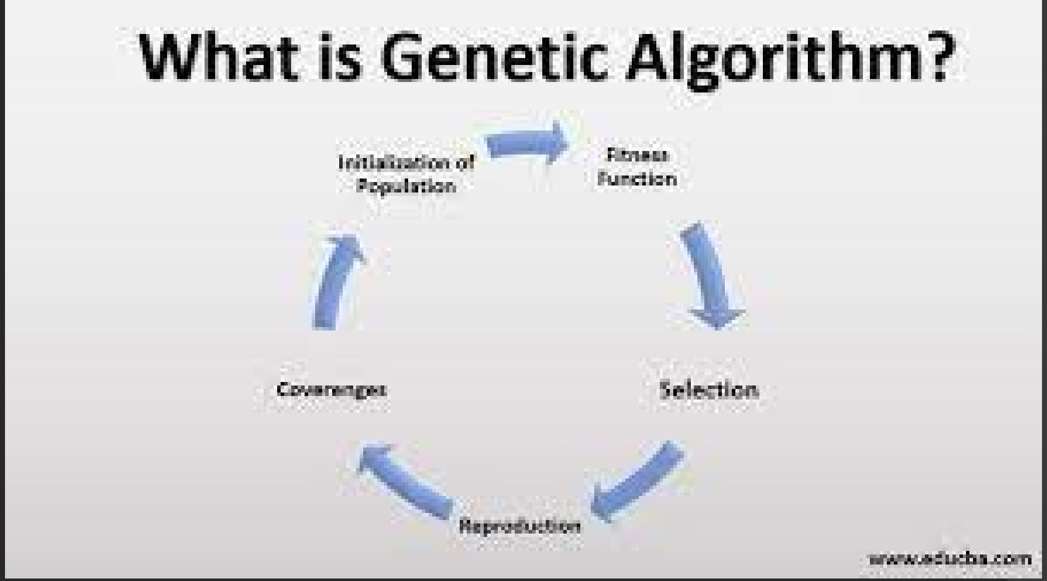
Feature selection’s aim is to filter out unrepresentative features from a given dataset. As there are no generally agreed financial ratios for bankruptcy prediction and credit scoring, collected variables must be examined for their representativeness, i.e., importance and explanatory power, in the chosen dataset.

Therefore, the performance of classifiers after performing feature selection could be enhanced over that of classifiers without feature selection.



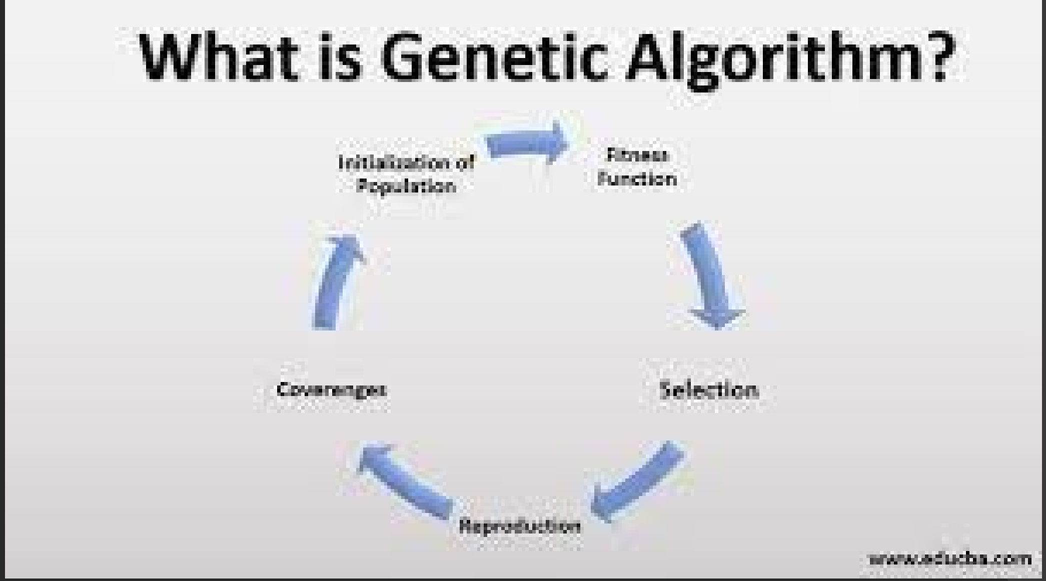
Genetic Algorithm (GA)

In Genetic algorithms (GA) a population of strings (called chromosomes), which encode candidate solutions (called individuals) to an optimization problem, evolves for better solutions. In general, the genetic information (i.e., chromosome) is represented by a bit string (such as binary strings of 0s and 1s) and sets of bits encode the solution.



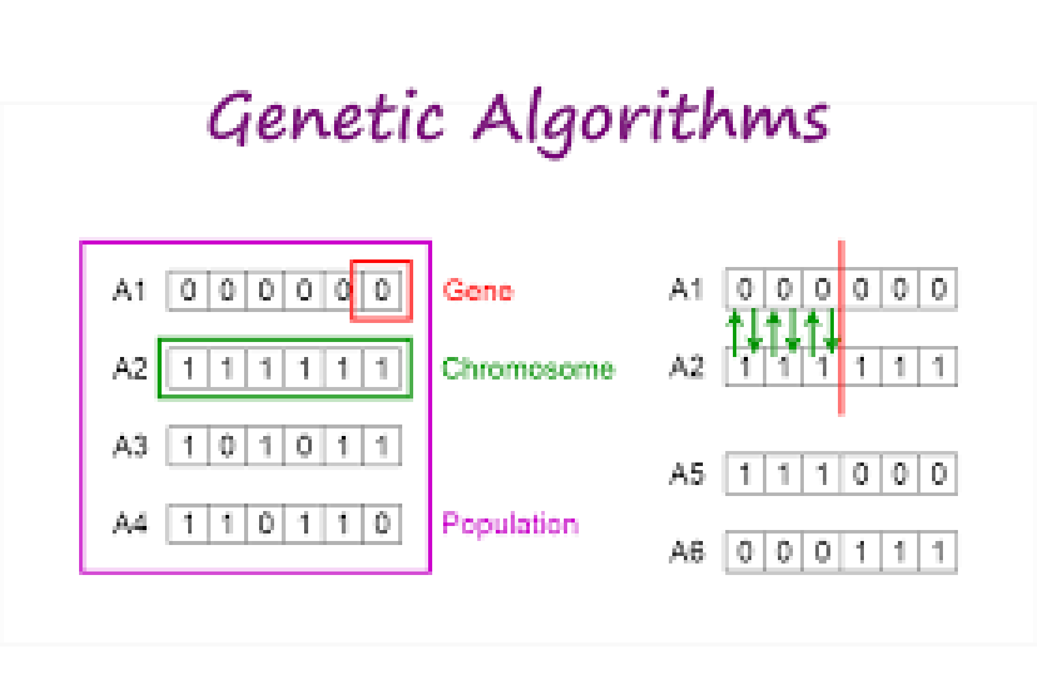
Genetic Algorithm (GA)

Then, genetic operators are applied to the individuals of the population for the next generation (i.e., a new population of individuals). There are two main genetic operators, which are crossover and mutation. Crossover creates two offspring strings from two parent strings copying selected bits from each parent.



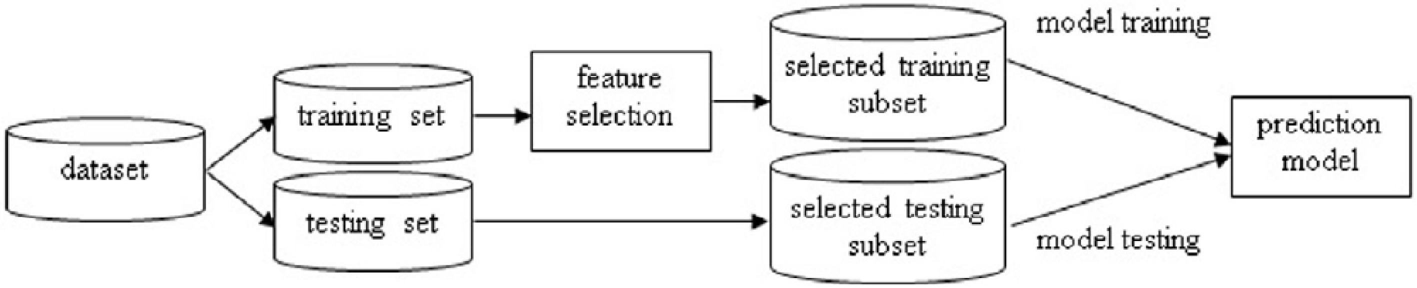
Genetic Algorithm (GA)

On the other hand, mutation randomly changes the value of a single bit (with small probability) to the bit strings. Furthermore, a fitness function is used to measure the quality of an individual in order to increase the probability that the single bit can survive throughout the evolutionary process.



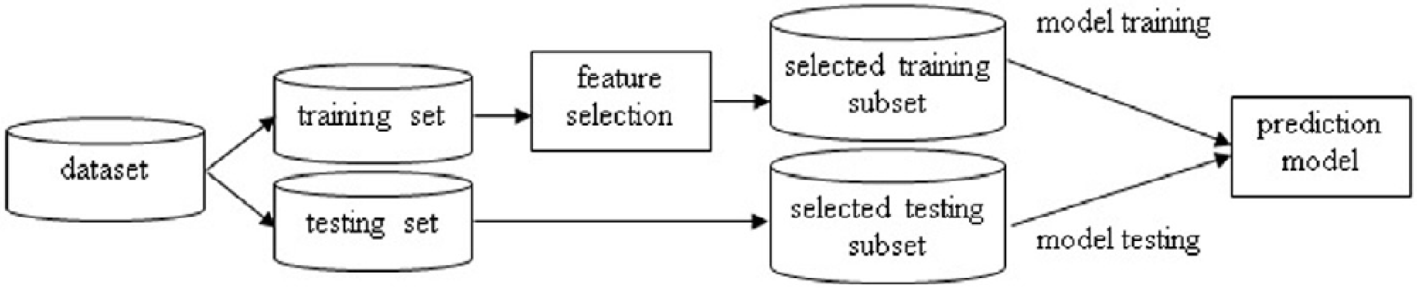
Filter based Feature selection for bankruptcy prediction

This figure outlines process of performing filter-based feature selection for financial distress prediction. The first step is to divide each dataset into the training and testing sets by 10-fold cross validation. Then, each feature selection method is executed over the training set. Next, the selected features are used as the new training set.



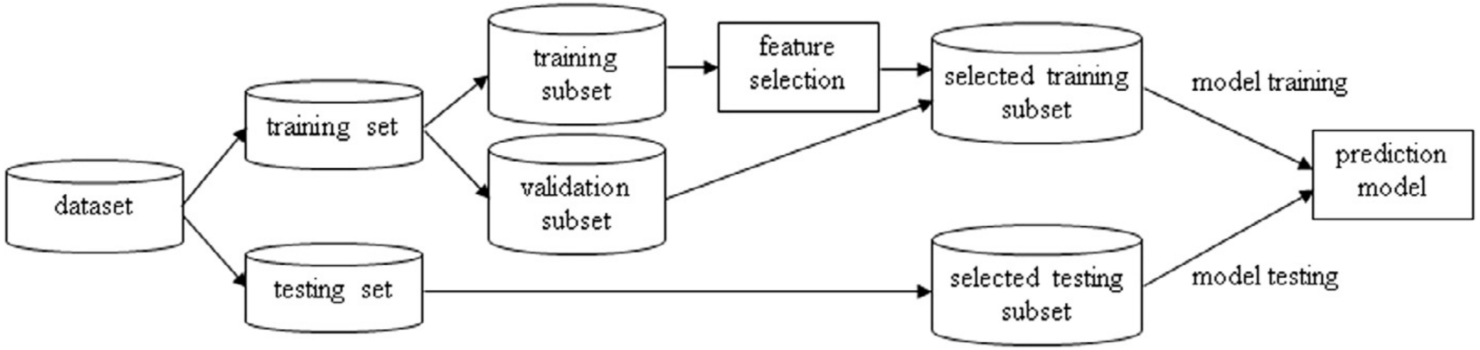
Filter based Feature selection for bankruptcy prediction

Finally, the testing set containing the same selected features as the new training set is used to test the performance of the prediction model. Note that the threshold to determine representative features by the filter-based feature selection methods is based on the feature, which is significant at the 0.05 level. For example, using the t-test method the features having the p values less than 0.05 are kept; otherwise they are filtered out.



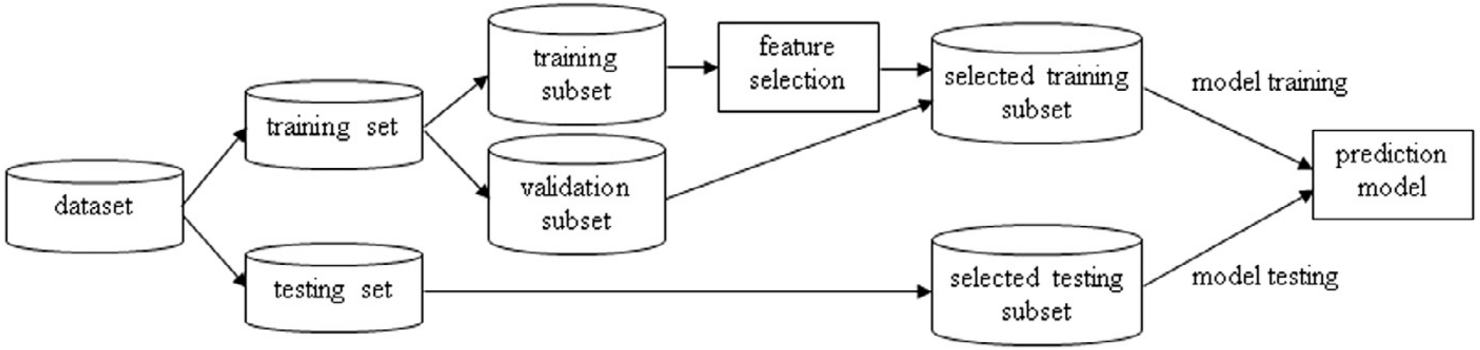
Wrapper based Feature selection for bankruptcy prediction

This figure shows the process of performing wrapper-based feature selection for financial distress prediction. First, each dataset is divided into the training and testing sets by 10-fold cross validation. Each training set is further sampled for the training and validation subsets to train the wrapper-based feature selection methods.



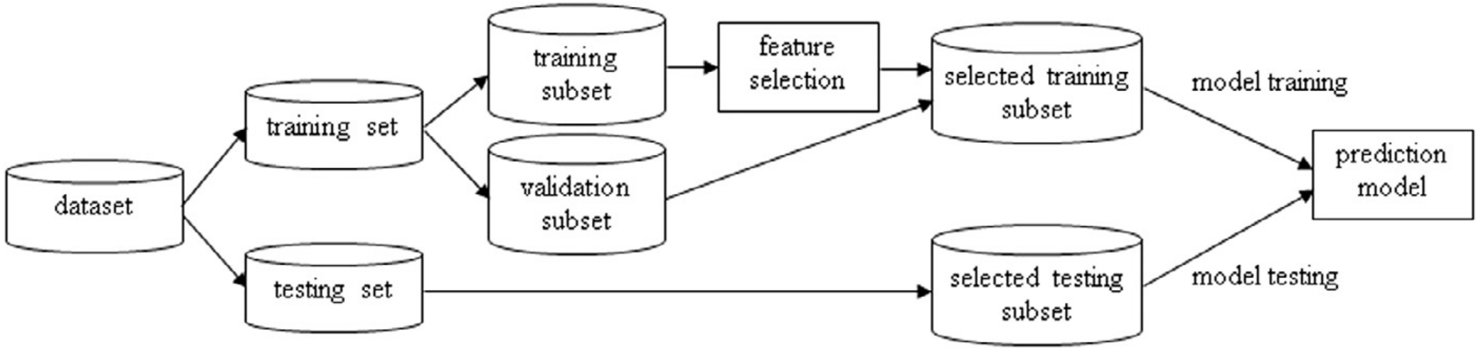
Wrapper based Feature selection for bankruptcy prediction

Then, the population pool is initialized where each group of the chromosome or particle represents the selected feature set. Next, each chromosome or particle in the population pool (as the training subset) is used to construct multiple models. After the models are constructed, the validation subset is used to test their accuracy.



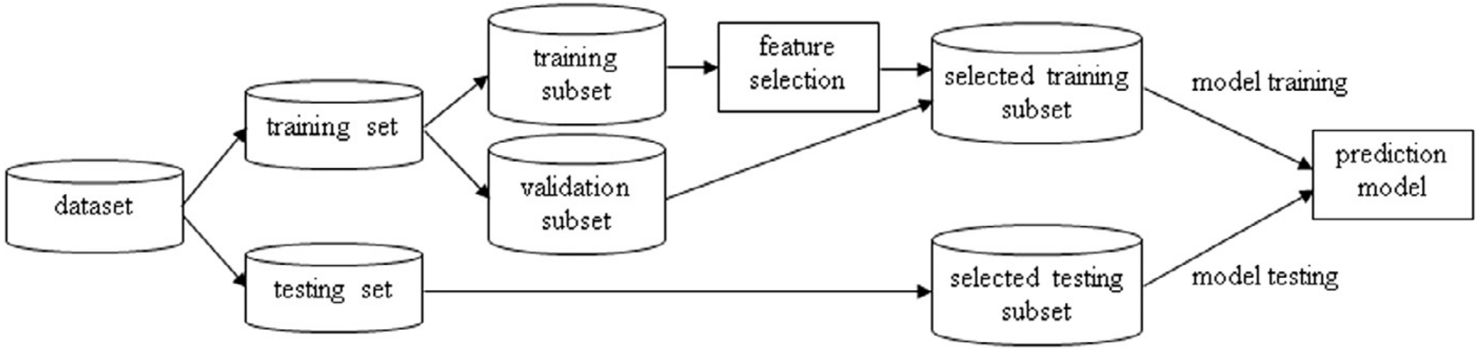
Wrapper based Feature selection for bankruptcy prediction

For GA, the performance of the models constructed by each chromosome is examined, and then the selection, crossover, mutation operations are performed to replace the current population pool. On the other hand, each particle of PSO is examined for its performance and its position and velocity in the feature space are adjusted by the sigmoid function to replace the current population pool.



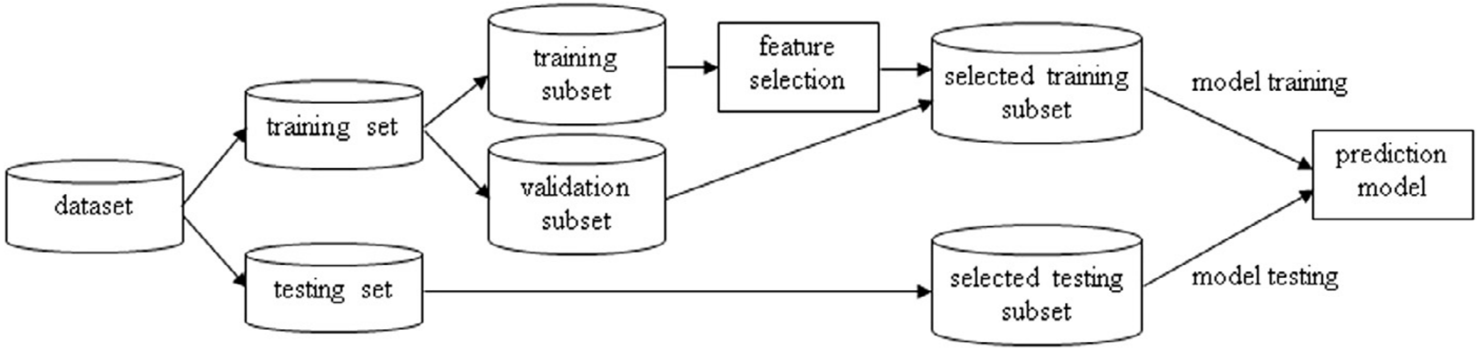
Wrapper based Feature selection for bankruptcy prediction

Consequently, the evolutionary process will be terminated until the stopping criterion is met. Then, the chromosome or particle having the highest accuracy over the validation subset is used as the training set to train the prediction model.



Wrapper based Feature selection for bankruptcy prediction

Finally, the testing set containing the same selected features as the chromosome or particle is used to test the performance of the prediction model.



Conclusion

However, on average GA performs better than the others over the credit scoring datasets whereas LR outperforms the other methods over the bankruptcy prediction datasets. Despite these findings, several feature selection methods have shown some promising results for bankruptcy prediction and credit scoring. In particular, t-test and LR as the filter methods and GA as the wrapper method can be used in the future.

Conclusion

It should be noted that performing feature selection does not always improve the models' performances, especially for CART and SVM. This may be because when constructing these models, CART can determine important features like many feature selection methods do during the tree construction process whereas SVM generally assigns some weights to the input features (i.e. attributes).

Conclusion

Moreover, since related studies, e.g. Clarke et al. [9], have shown the advantage of SVM for high dimensional data, the dimensionalities of financial distress datasets are relatively small compared with other domains, such as genomic and proteomic problems. Therefore, the need of performing the feature selection step for credit scoring and bankruptcy prediction depends on the chosen classifiers.

Future.Works

The chosen filter and wrapper based feature selection methods are based on the mostly used methods in bankruptcy prediction and credit scoring, other filter (e.g. information gain) and wrapper (e.g. naïve Bayes) methods can also be employed for the feature selection task.

In addition to using single classification techniques to develop the prediction models, combining multiple classifiers or classifier ensembles by the bagging and boosting combination methods can be developed.

Thank You for your attention.

Q&A