https://archive.ics.uci.edu/ml/datasets/SECOM

|  |  |
| --- | --- |
| **SECOM Data Set**  *Download*: [Data Folder](https://archive.ics.uci.edu/ml/machine-learning-databases/secom/), [Data Set Description](https://archive.ics.uci.edu/ml/machine-learning-databases/secom/secom.names)  **Abstract**: Data from a semi-conductor manufacturing process | https://archive.ics.uci.edu/ml/assets/MLimages/Large179.jpg |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 1567 | **Area:** | Computer |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 591 | **Date Donated** | 2008-11-19 |
| **Associated Tasks:** | Classification, Causal-Discovery | **Missing Values?** | Yes | **Number of Web Hits:** | 142769 |

**Source:**

Authors: Michael McCann, Adrian Johnston

**Data Set Information:**

A complex modern semi-conductor manufacturing process is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. It is often the case that useful information is buried in the latter two. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be applied to identify the most relevant signals. The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs.   
  
To enhance current business improvement techniques the application of feature selection as an intelligent systems technique is being investigated.   
  
The dataset presented in this case represents a selection of such features where each example represents a single production entity with associated measured features and the labels represent a simple pass/fail yield for in house line testing, figure 2, and associated date time stamp. Where –1 corresponds to a pass and 1 corresponds to a fail and the data time stamp is for that specific test point.   
  
  
Using feature selection techniques it is desired to rank features according to their impact on the overall yield for the product, causal relationships may also be considered with a view to identifying the key features.   
  
Results may be submitted in terms of feature relevance for predictability using error rates as our evaluation metrics. It is suggested that cross validation be applied to generate these results. Some baseline results are shown below for basic feature selection techniques using a simple kernel ridge classifier and 10 fold cross validation.   
  
Baseline Results: Pre-processing objects were applied to the dataset simply to standardize the data and remove the constant features and then a number of different feature selection objects selecting 40 highest ranked features were applied with a simple classifier to achieve some initial results. 10 fold cross validation was used and the balanced error rate (\*BER) generated as our initial performance metric to help investigate this dataset.   
  
  
SECOM Dataset: 1567 examples 591 features, 104 fails   
  
FSmethod (40 features) BER % True + % True - %   
S2N (signal to noise) 34.5 +-2.6 57.8 +-5.3 73.1 +2.1   
Ttest 33.7 +-2.1 59.6 +-4.7 73.0 +-1.8   
Relief 40.1 +-2.8 48.3 +-5.9 71.6 +-3.2   
Pearson 34.1 +-2.0 57.4 +-4.3 74.4 +-4.9   
Ftest 33.5 +-2.2 59.1 +-4.8 73.8 +-1.8   
Gram Schmidt 35.6 +-2.4 51.2 +-11.8 77.5 +-2.3

**Attribute Information:**

Key facts: Data Structure: The data consists of 2 files the dataset file SECOM consisting of 1567 examples each with 591 features a 1567 x 591 matrix and a labels file containing the classifications and date time stamp for each example.   
  
As with any real life data situations this data contains null values varying in intensity depending on the individuals features. This needs to be taken into consideration when investigating the data either through pre-processing or within the technique applied.   
  
The data is represented in a raw text file each line representing an individual example and the features seperated by spaces. The null values are represented by the 'NaN' value as per MatLab.