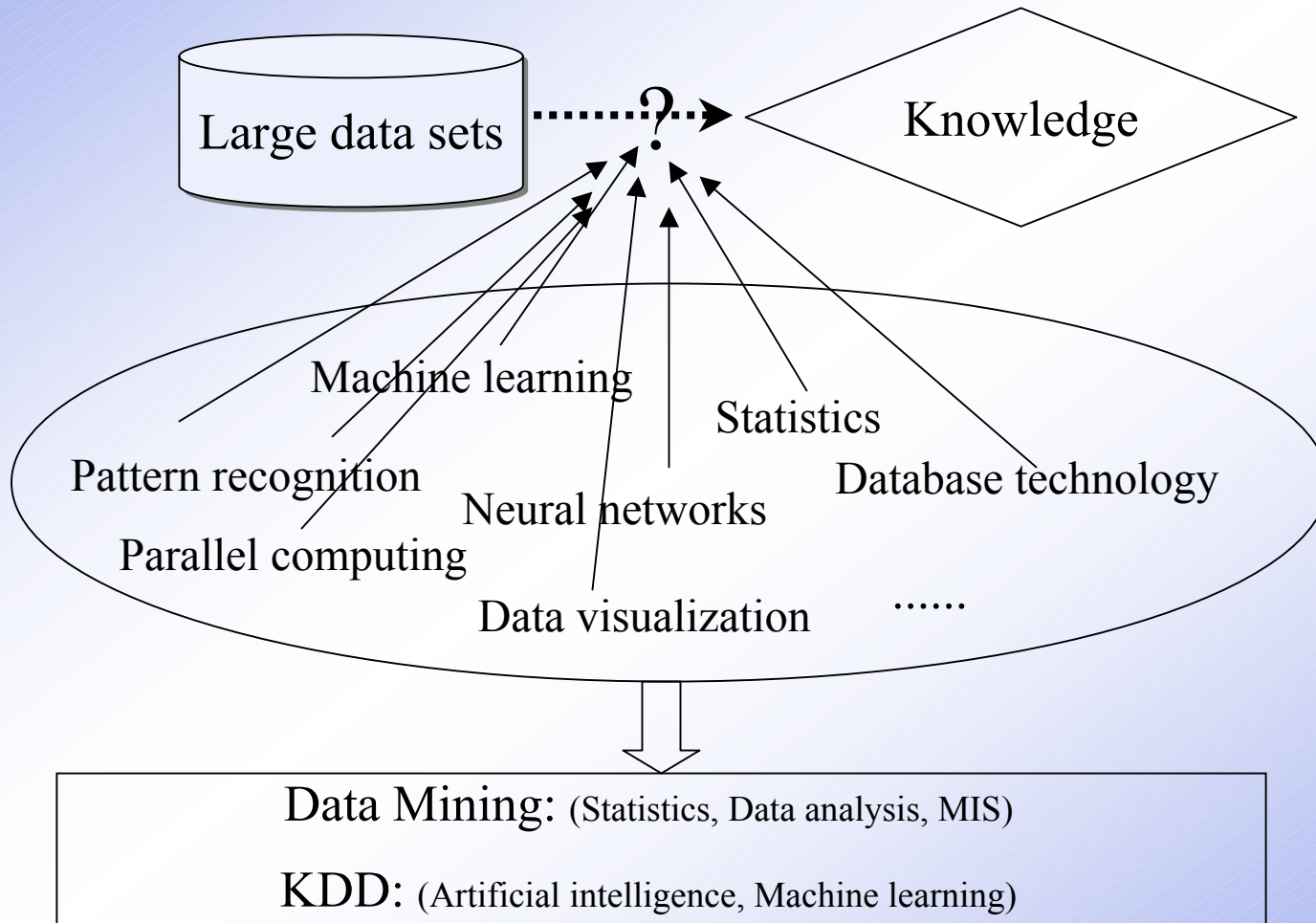
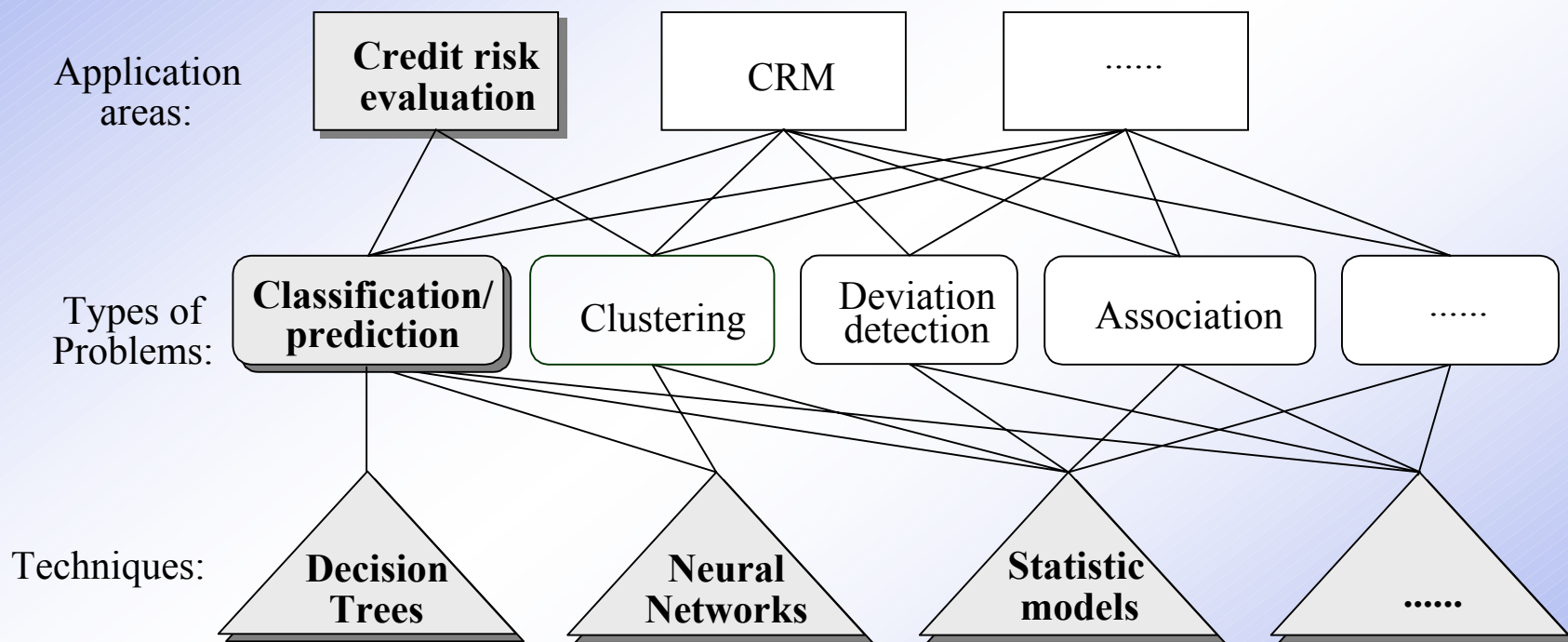
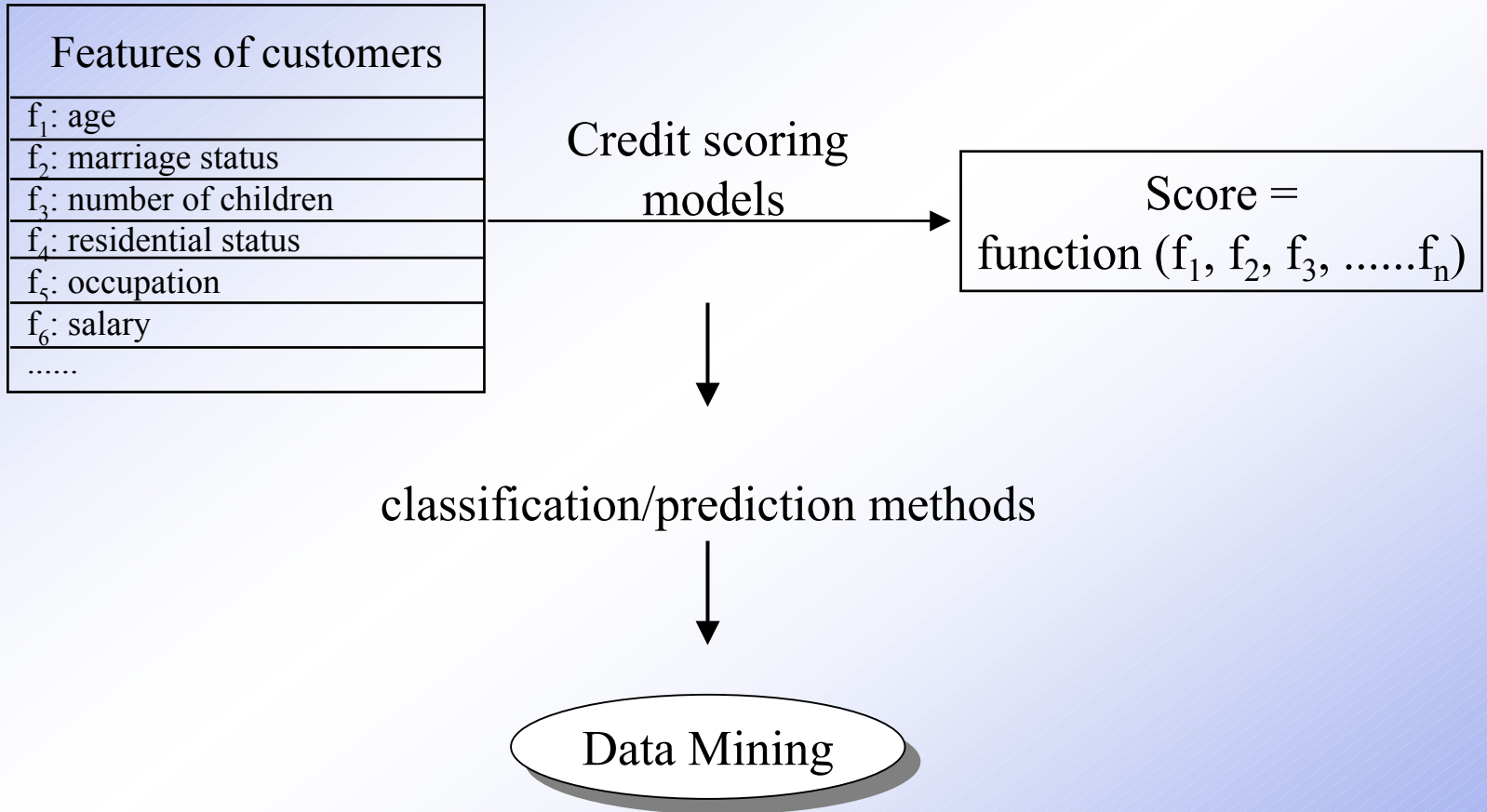


Data Mining Feature Selection for Credit Scoring Models

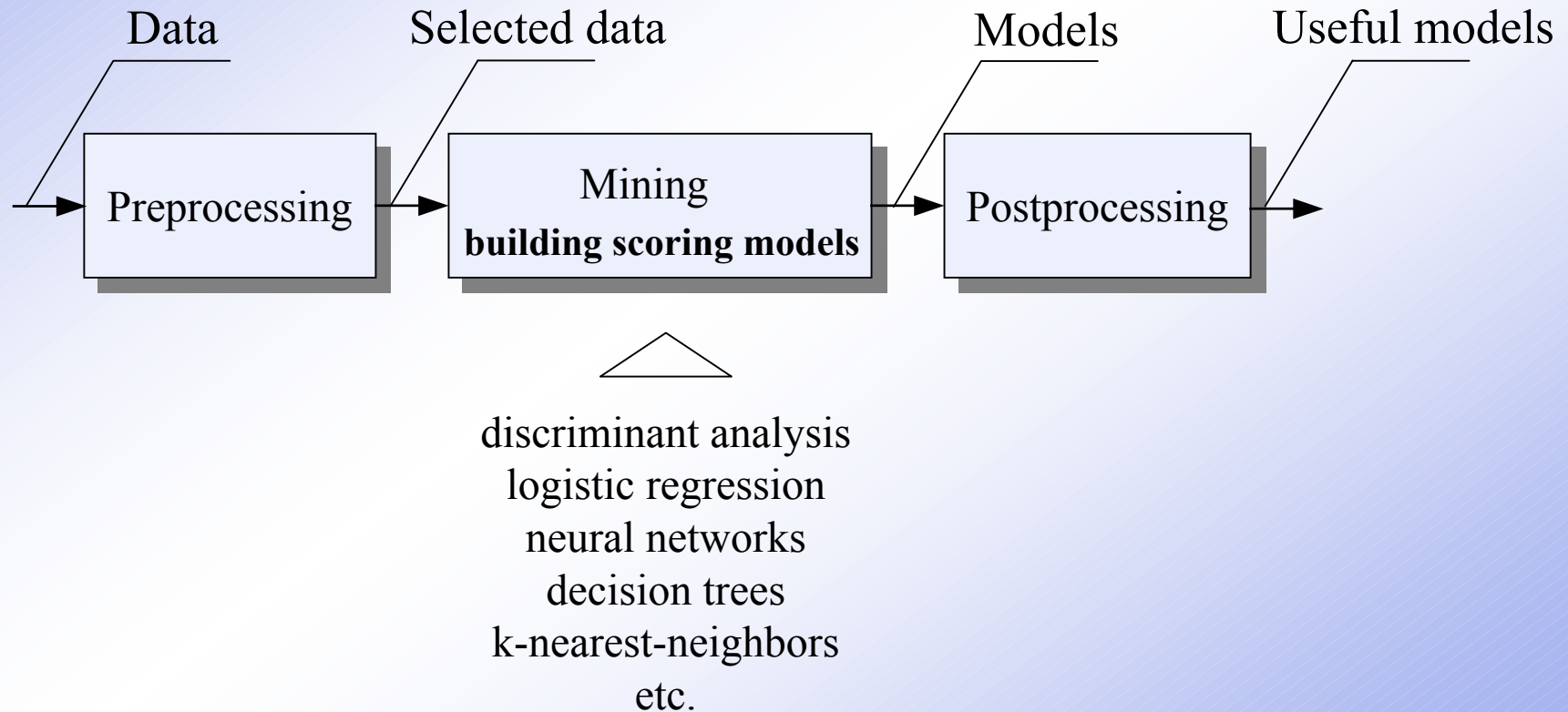
- Background introduction
- The used methods and algorithms
- The process of the study
- Results of the study



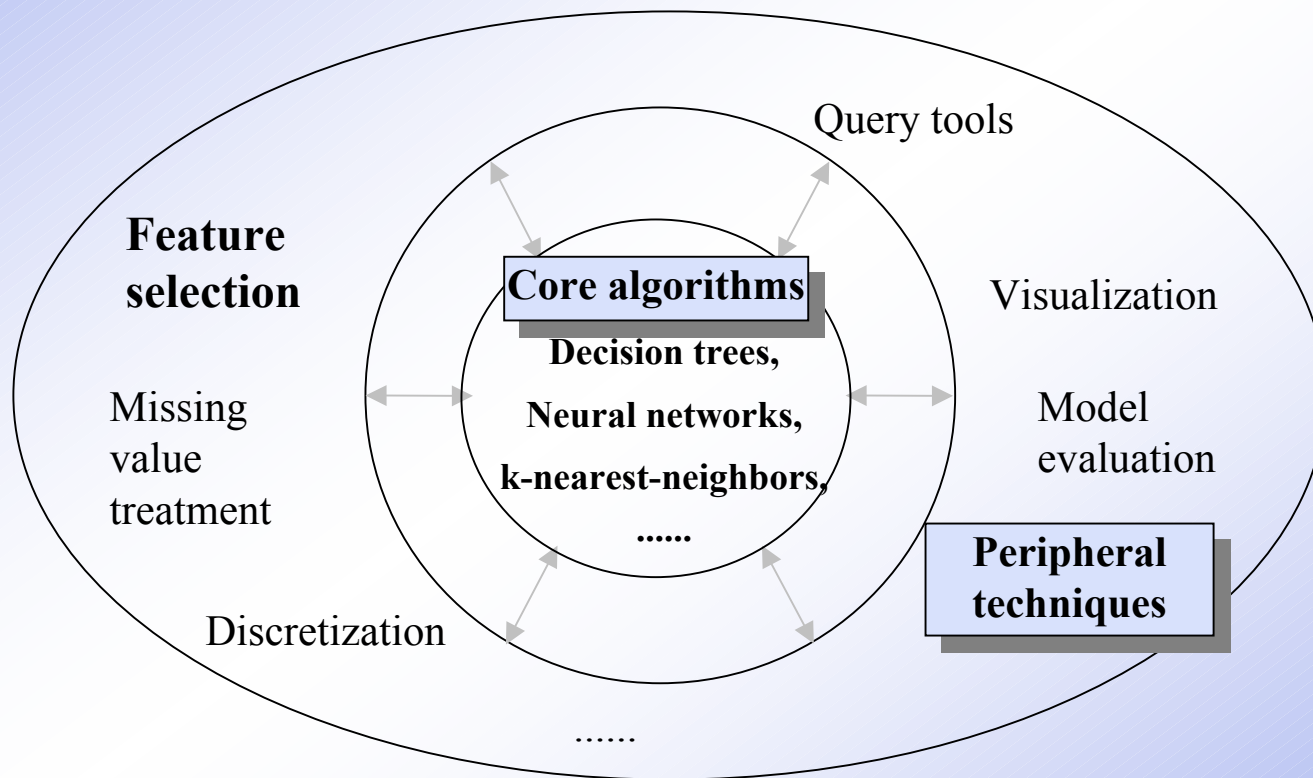




Process of Knowledge Discovery/Data Mining:



The components of data mining techniques



Feature selection in credit scoring: WHY

Irrelevant, redundant features

- more data, longer time
- misleading or overfitting
- complex model



A good choice of features

- less data, faster
- higher accuracy
- simpler model

Feature selection in credit scoring: HOW

Unsystematic process:

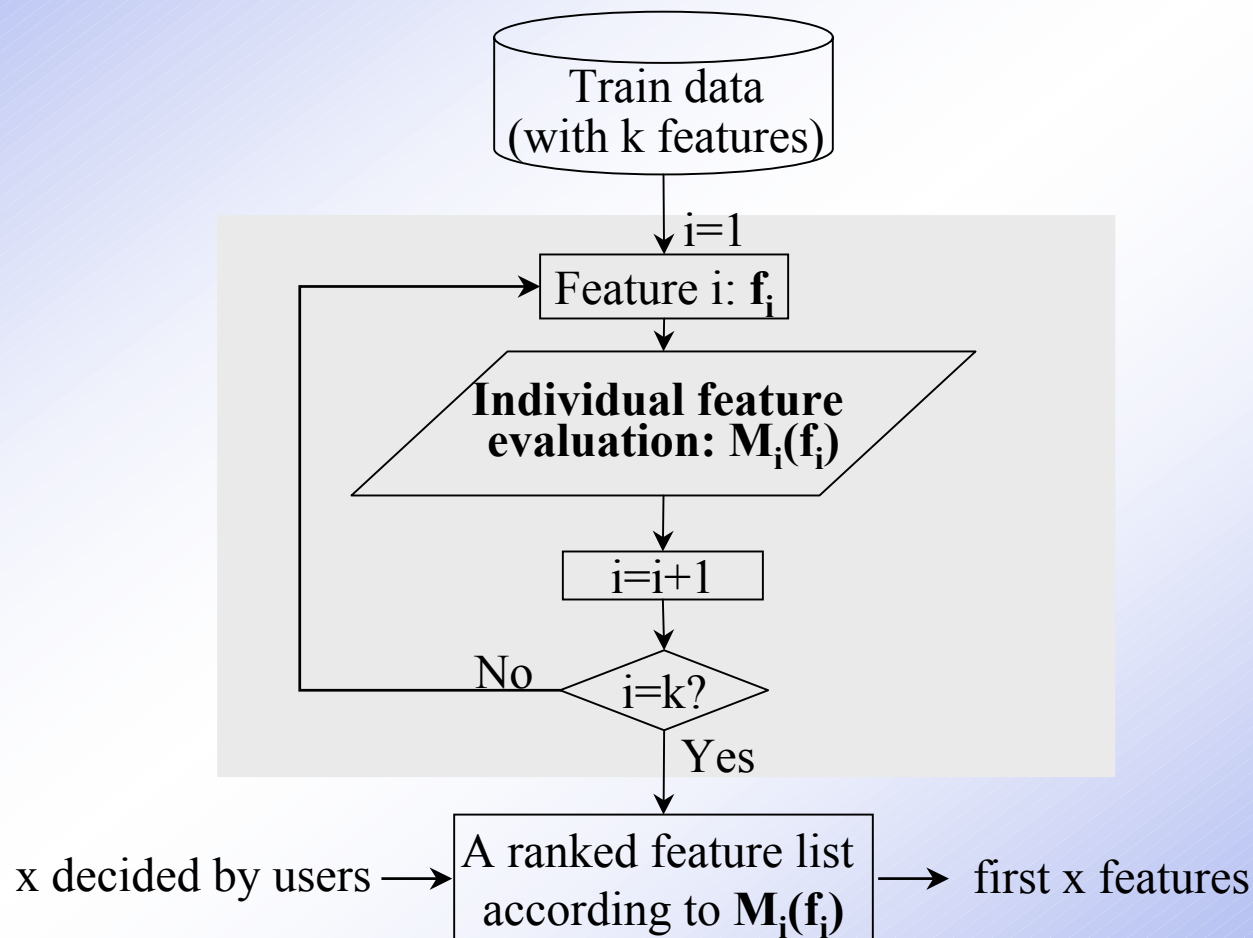
- . Identify important variables
- . Eliminate correlated variables
- . Select features during the model building
 - e. g. stepwise statistical procedures
- . Expert knowledge and experience
-

Feature selection in credit scoring: HOW

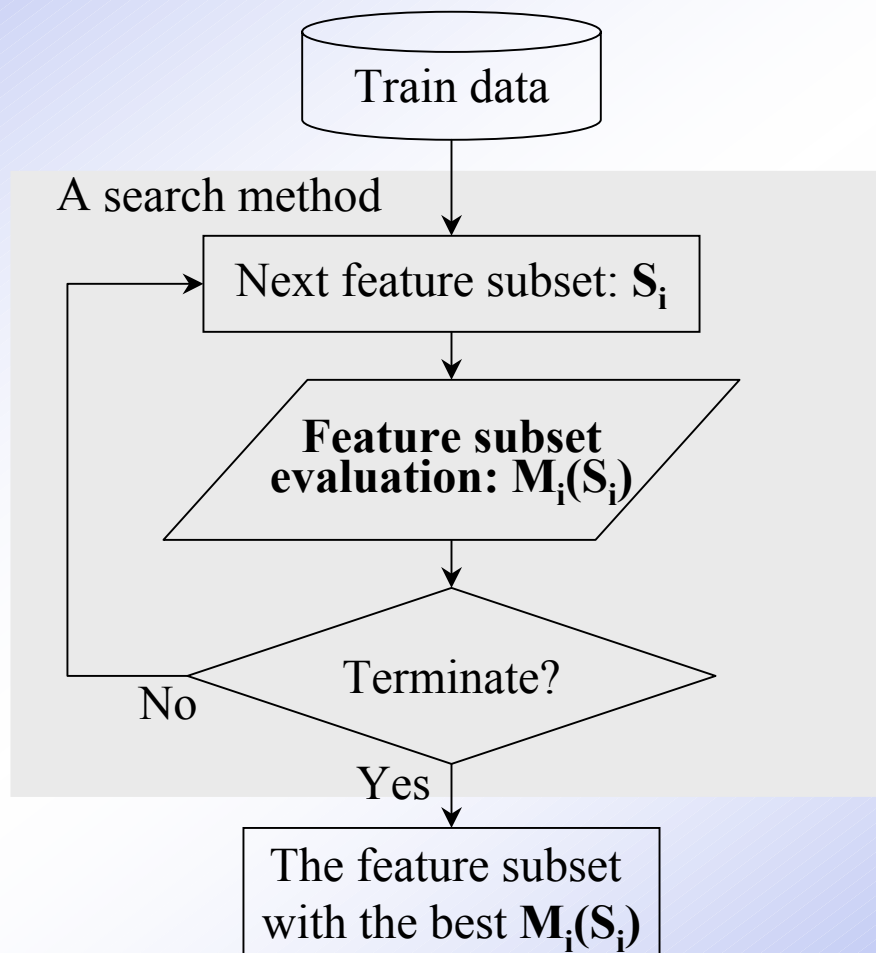
Formal process:

Data mining feature selection
using machine learning algorithms

Data mining feature selection methods: Feature ranking algorithms

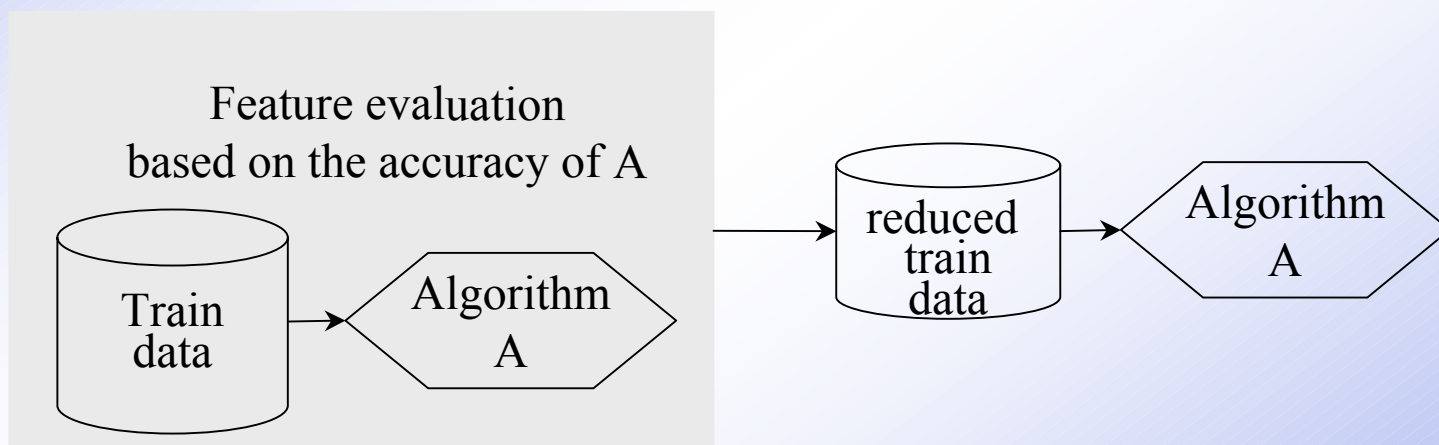


Data mining feature selection methods: Best feature subset algorithms



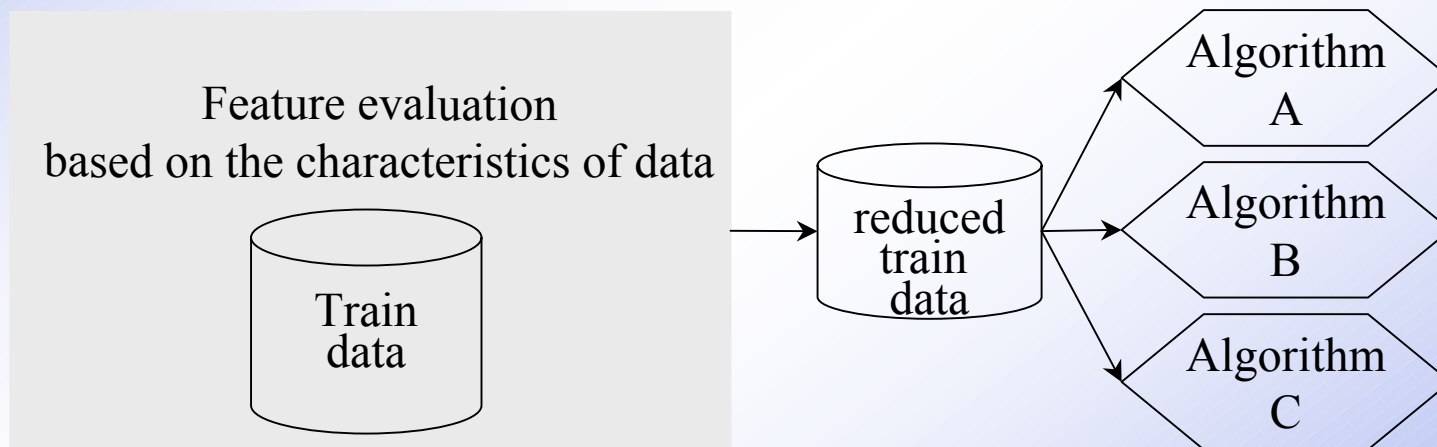
Data mining feature selection methods: Wrapper algorithms

Wrapper Algorithms



Data mining feature selection methods: Filter algorithms

Filter Algorithms



The used four feature selection methods

	Filter Algorithm	Wrapper Algorithm
Feature ranking algorithm	ReliefF (REF)	
Best feature subset algorithm	Correlation-based (CFS) Consistency-based (CON)	Wrapper(WRP)

Evaluation measure of the ReliefF method (REF)

set $M_{REF}(A) = 0$;

for $i=1$ to m do (m is a user specified number):

Begin

 randomly sample an example R from the dataset;

 find its nearest neighbor from the same class (example S);

 find its nearest neighbor from the different class (example D);

$M_{REF}(A) = M_{REF}(A) - \text{diff}(A, R, S)/m + \text{diff}(A, R, D)/m$;

end;

Evaluation measure of the correlation-based method (CFS)

$$M_{\text{CFS}} = \frac{\overline{kr_{\text{cf}}}}{\sqrt{k + k(k-1)\overline{r_{\text{ff}}}}}$$

r_{cf} : feature-class correlation

r_{ff} : feature- feature correlation

Evaluation measure of the consistency-based method (CON)

$$M_{\text{CON}} = \text{consistency rate} = 1 - \text{inconsistency rate}$$

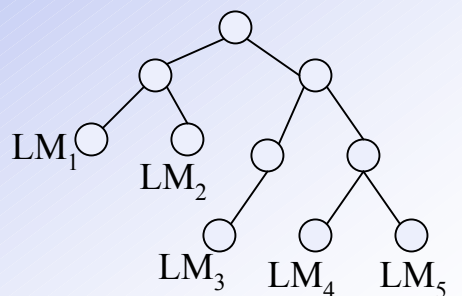
Example of inconsistent patterns:

A (1, 0, 0, C_1) and B(1, 0, 0, C_2)

Evaluation measure of the wrapper method (WRP)

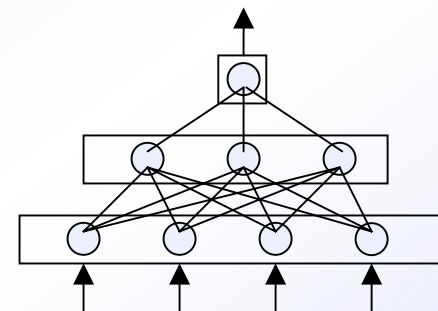
$$\mathbf{M}_{\text{WRP}} = \text{classification accuracy} = 1 - \text{test error rate}$$

The used three classification learning methods

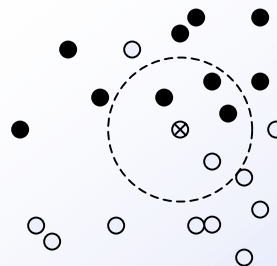


LM_i : linear regression models

Model Tree (M5)



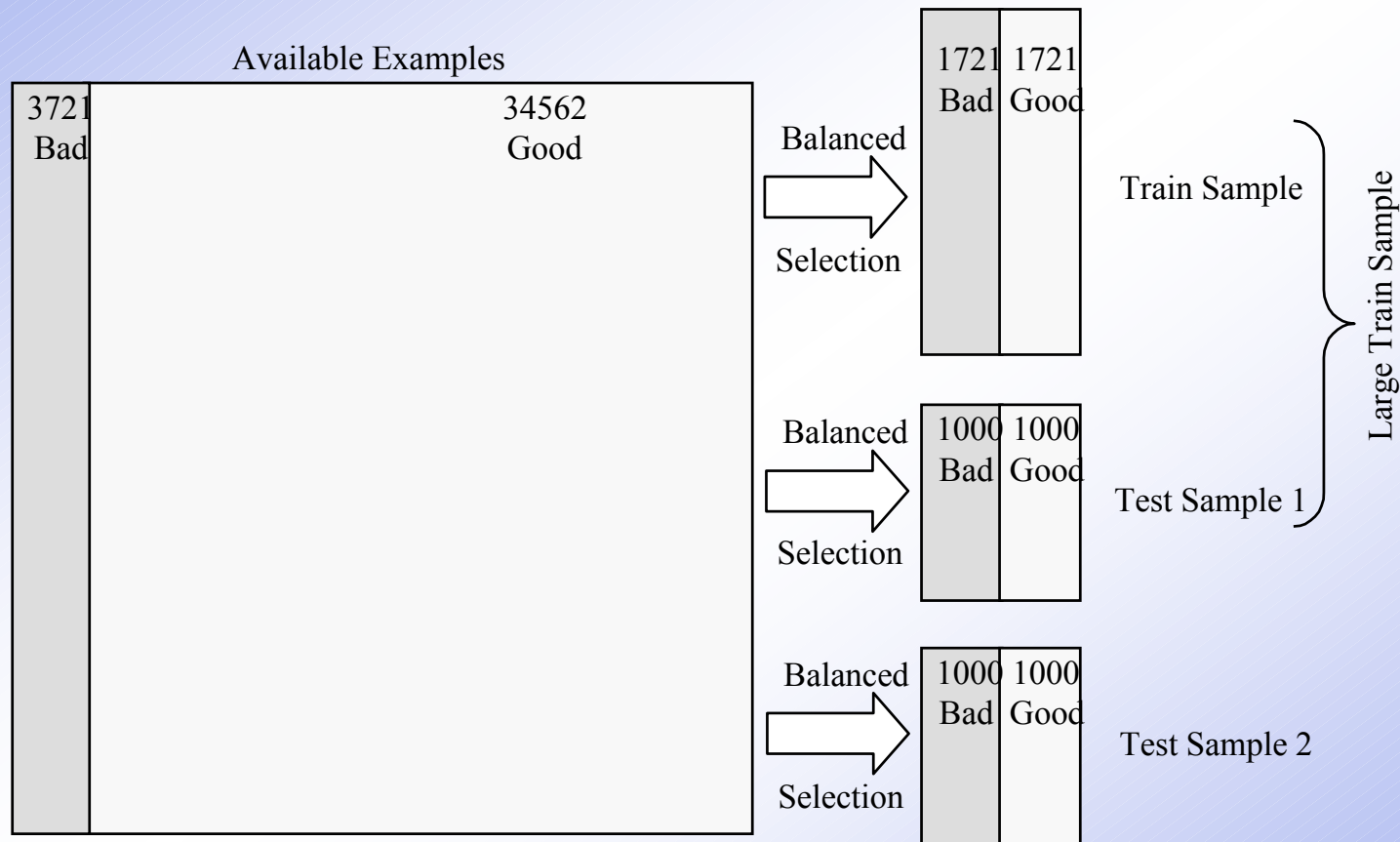
Multi-layer perceptron neural network (MLP)



k-Nearest-Neighbors (k-NN)

Credit information from bank (72 features):

- Information of companies' accounts
- Evaluations of companies' financial status
- Evaluations of companies' creditworthiness
-



The process of the feature selection study

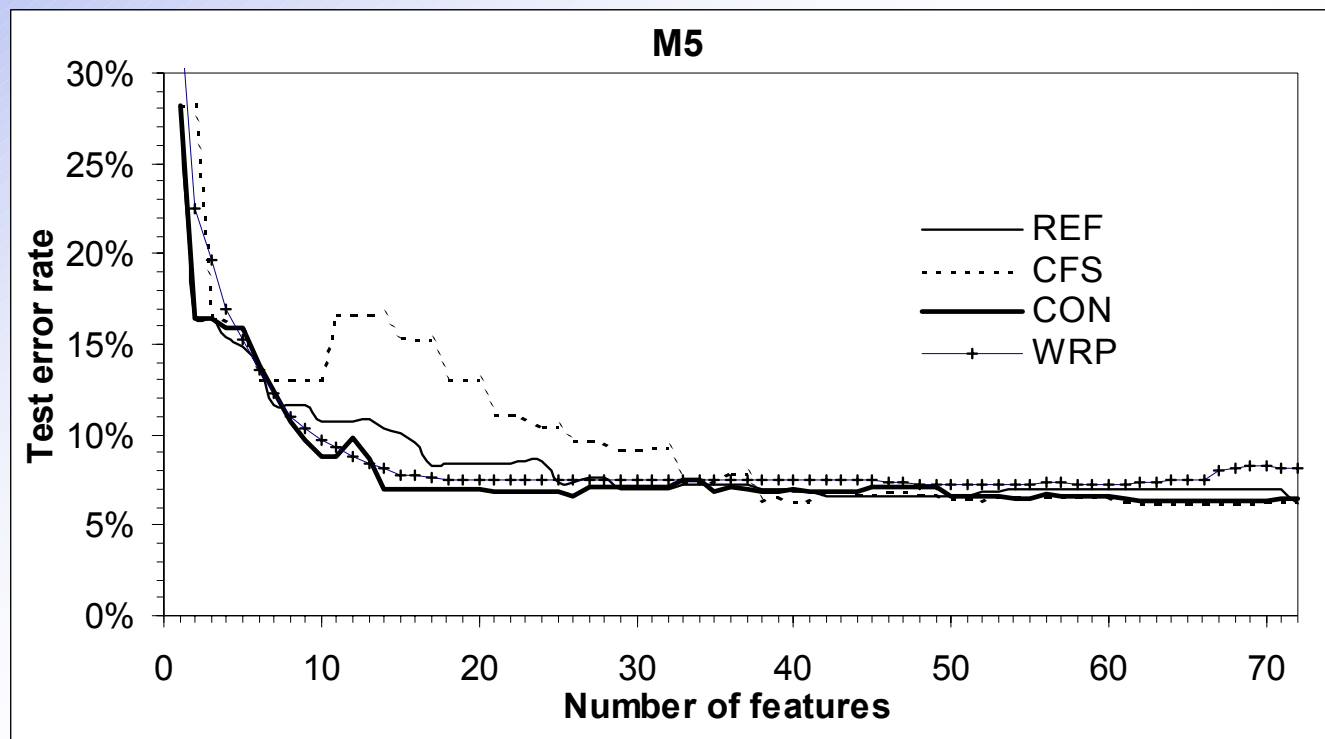
- ▶ 1. Ranking the features using four feature selection methods
2. Creating the learning curves for each algorithm
3. Selecting the used features for each algorithm
4. Training the final models with the selected features

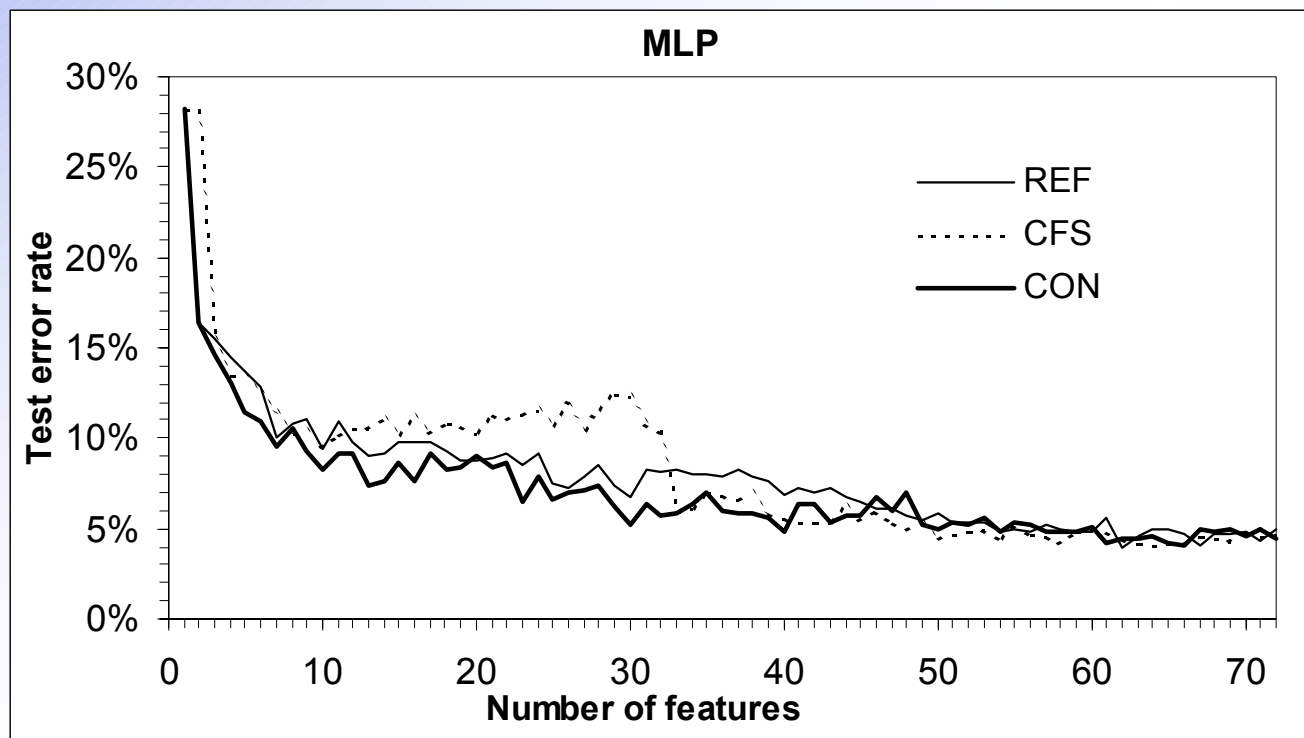
REF: feature ranking algorithm

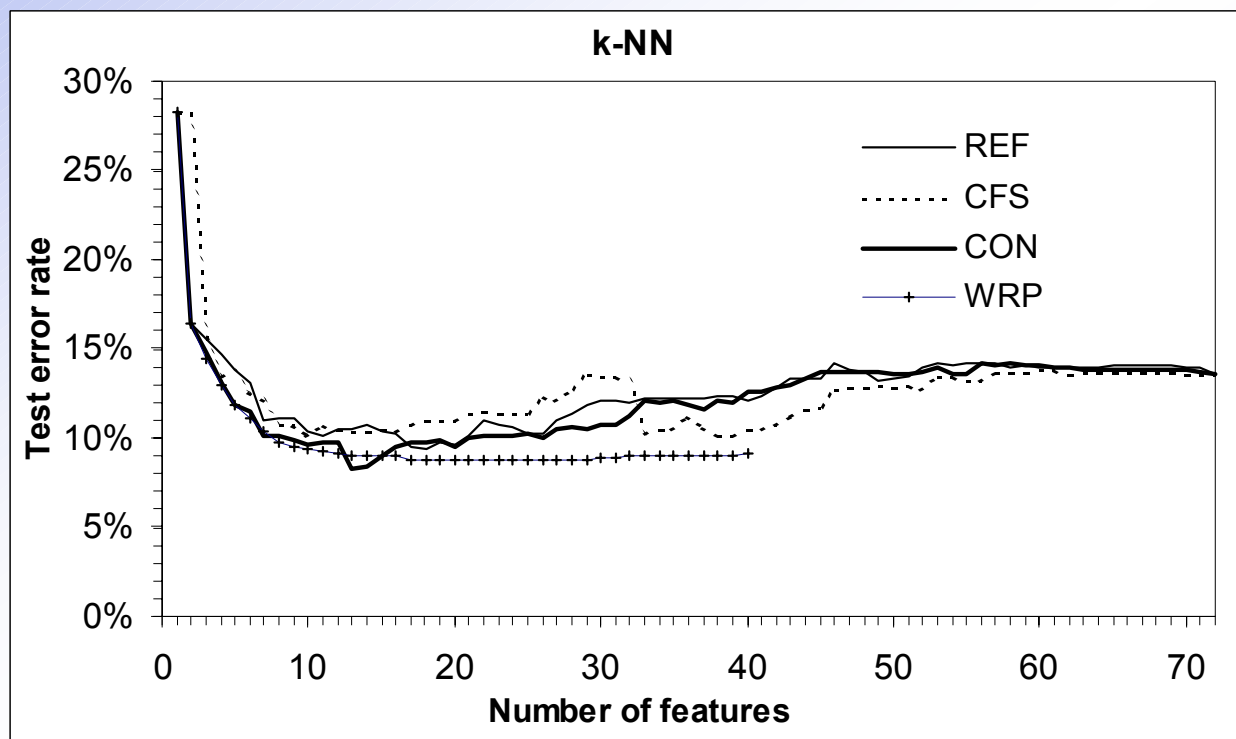
CFS: }
CON: } feature subset algorithms
WRP: } (forward hill climbing search,
features are selected one by one)

The process of the feature selection study


1. Ranking the features using different feature selection method
- ▶ 2. Creating the learning curves for each algorithm
3. Selecting the used features for each algorithm
4. Training the final models with the selected features

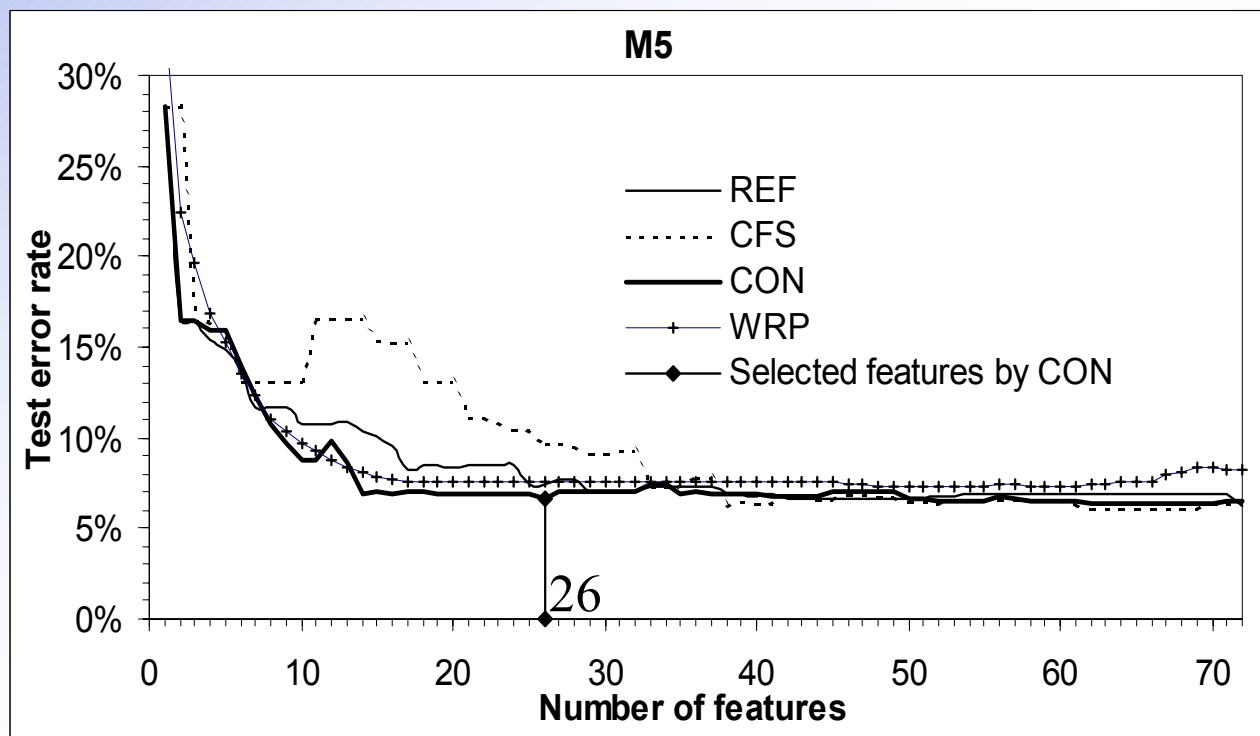


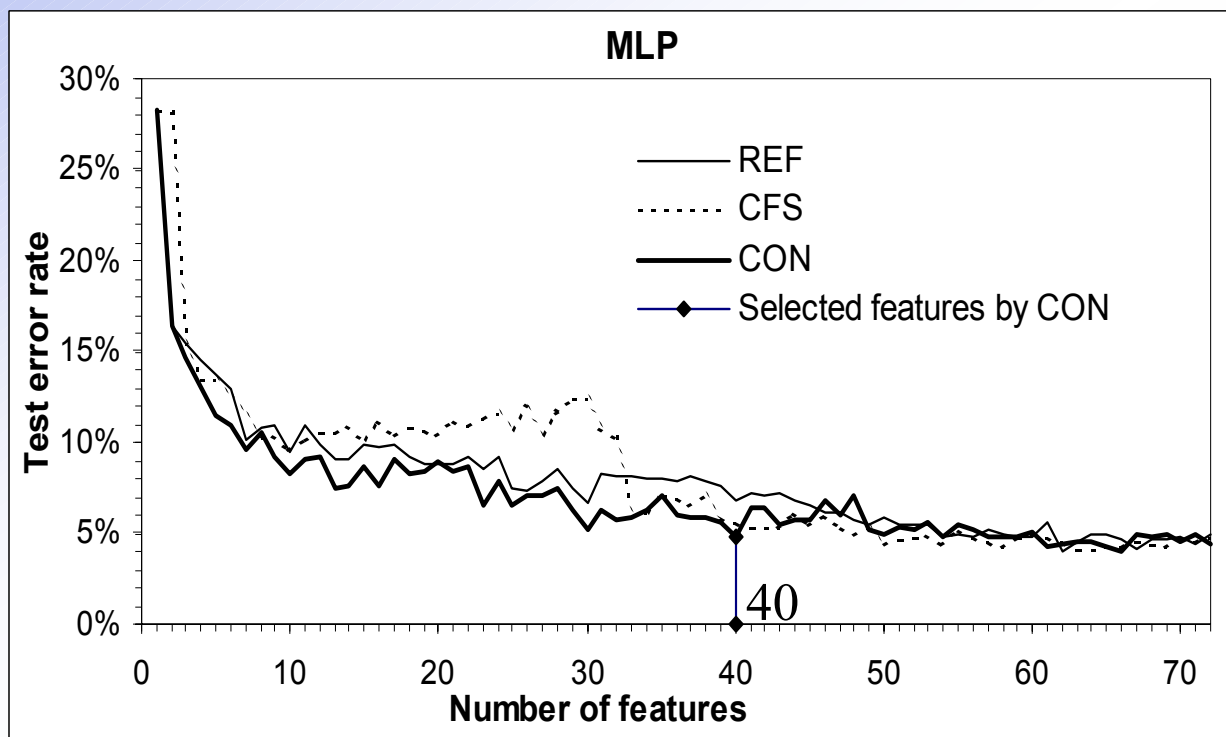


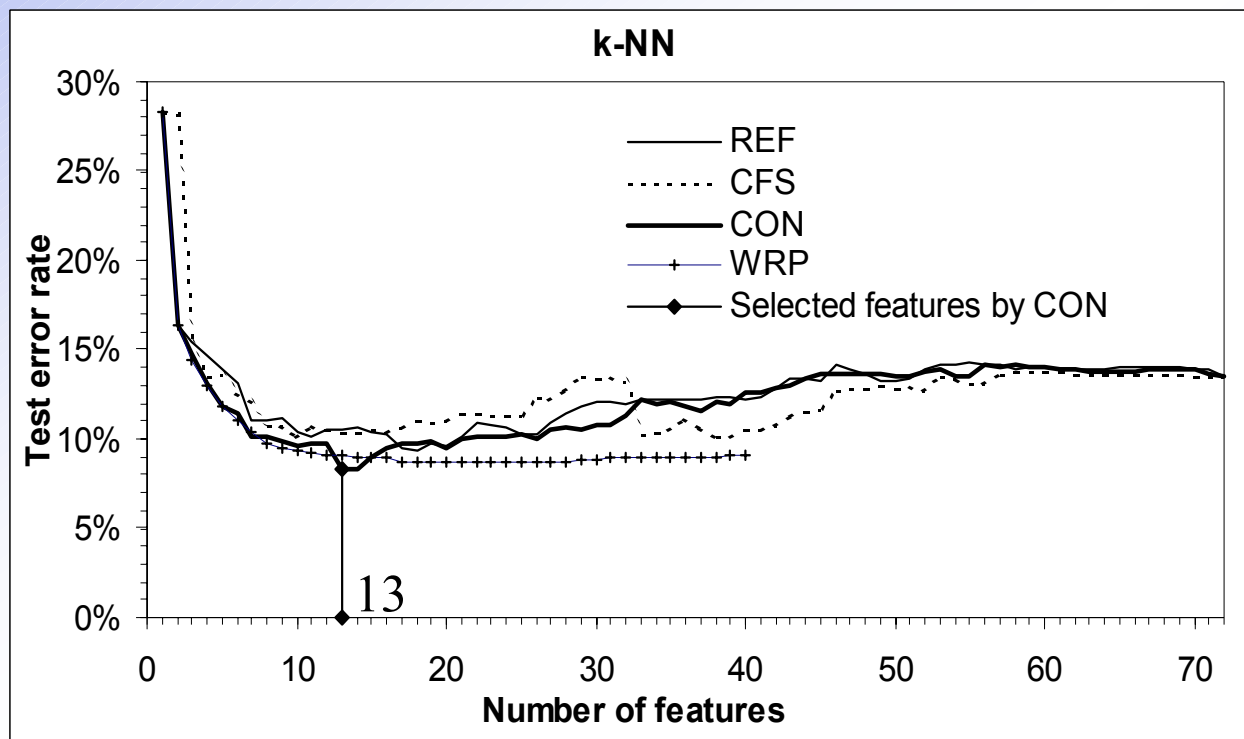


The process of the feature selection study

1. Ranking the features using different feature selection method
2. Creating the learning curves for each algorithm
-  3. Selecting the used features for each algorithm
4. Training the final models with the selected features







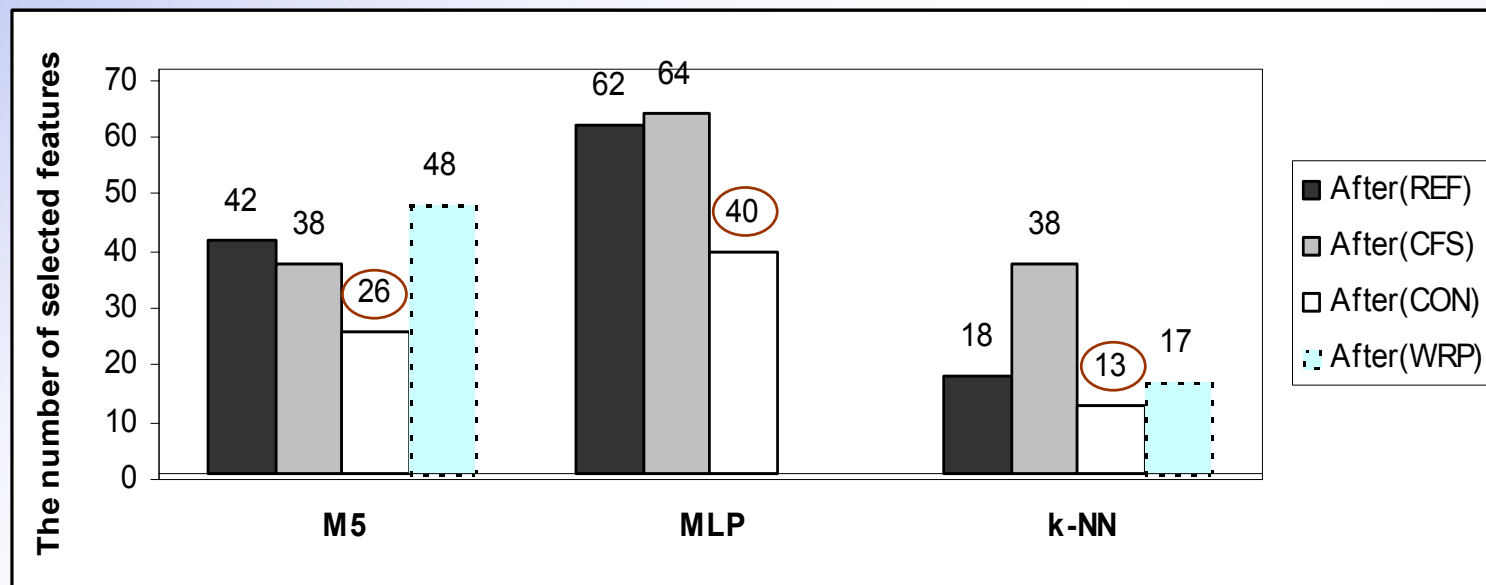
The process of the feature selection study

1. Ranking the features using different feature selection method
2. Creating the learning curves for each algorithm
3. Selecting the used features for each algorithm
- ▶ 4. Training the final models with the selected features

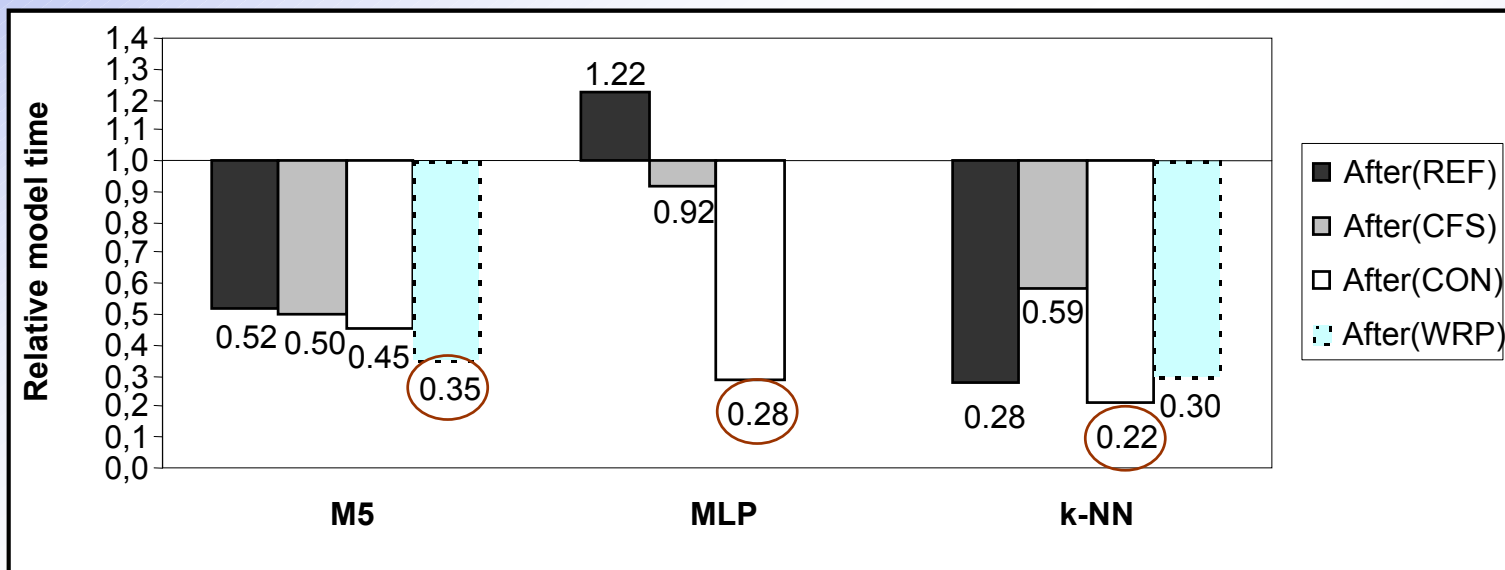
Performance of models before and after feature selection

1. Model Simplicity
2. Model Speed
3. Model Accuracy

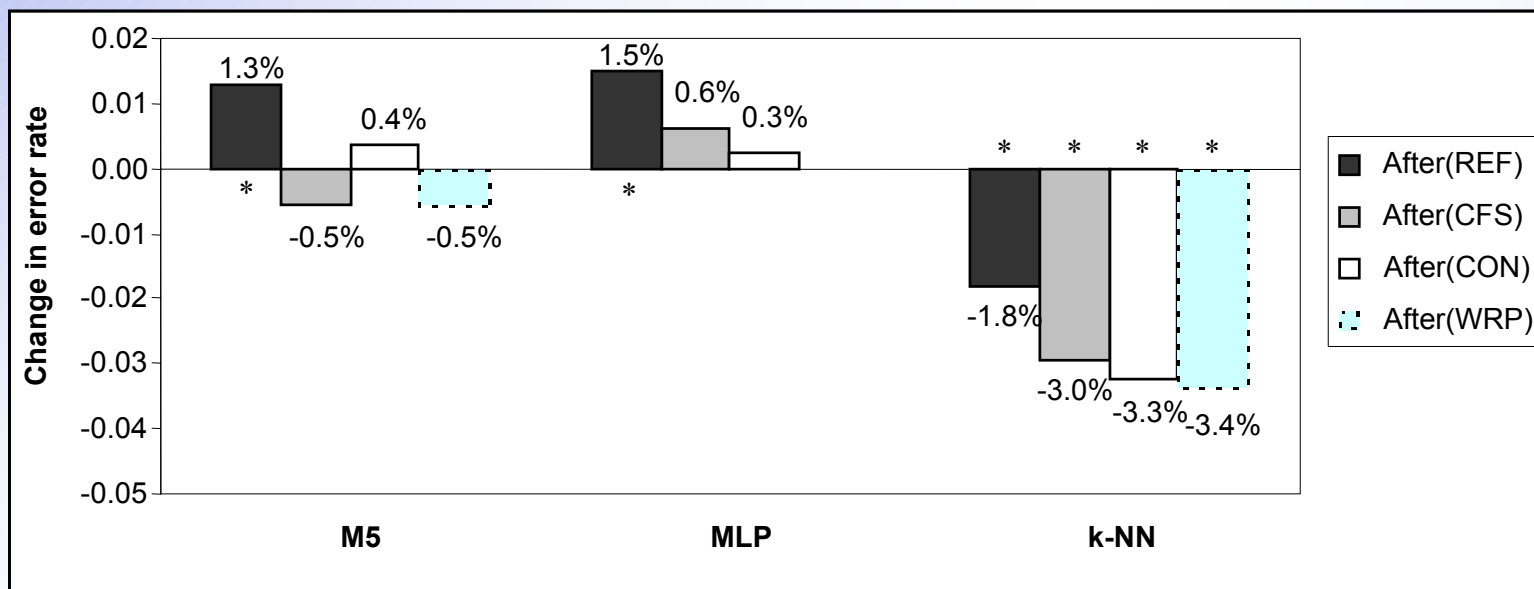
The number of selected features



Relative model time



Change in error rate



	REF	CFS	CON	WRP
M5	1). ++ 2). ++ 3). -	1). ++ 2). ++ 3). 0	1). +++ 2). ++ 3). 0	1). ++ 2). +++ 3). 0
MLP	1). + 2). - 3). -	1). + 2). + 3). 0	1). ++ 2). +++ 3). 0	
K-NN	1). +++ 2). +++ 3). +	1). ++ 2). ++ 3). +++	1). +++ 2). +++ 3). +++	1). +++ 2). +++ 3). +++

- 1). Simplicity
- 2). Speed
- 3). Accuracy

0: no change
-: worse
+: better

	REF	CFS	CON	WRP
M5	1). ++ 2). ++ 3). -	1). ++ 2). ++ 3). 0	1). +++ 2). ++ 3). 0	1). ++ 2). +++ 3). 0
MLP	1). + 2). - 3). -	1). + 2). + 3). 0	1). ++ 2). +++ 3). 0	
K-NN	1). +++ 2). +++ 3). +	1). ++ 2). ++ 3). +++	1). +++ 2). +++ 3). +++	1). +++ 2). +++ 3). +++

- 1). Simplicity
- 2). Speed
- 3). Accuracy

- 0: no change
-: worse
+: better

Data Mining Feature Selection for Credit Scoring Models

End