Extração e Seleção de Atributos

Ricardo Prudêncio

Introdução

- A *eficiência* de uma solução depende em muitos casos do tamanho do problema
- Dimensão de um problema de aprendizado
  - Número de atributos
  - Número de exemplos de treinamento
- A *dimensão* de um problema de aprendizado interfere em muitos casos na:
  - qualidade das respostas (*precisão*) dos algoritmos
  - e no *custo* do aprendizado

Redução de Dimensionalidade

- Tarefas:
  - Redução do número de atributos
  - Redução do número de exemplos

- Objetivos
  - Diminuir o custo do aprendizado
  - Aumentar o desempenho preditivo do algoritmo
  - Gerar modelos compactos mais fáceis de interpretar

Redução de Atributos

- Em geral, espera-se que todos os atributos sejam relevantes porém nem sempre é possível garantir isso
- Além disso, alguns atributos são redundantes e assim poderiam ser eliminados
- Objetivo:
  - Definir conjunto de atributos que sejam relevantes e não-redundantes

Redução de Atributos

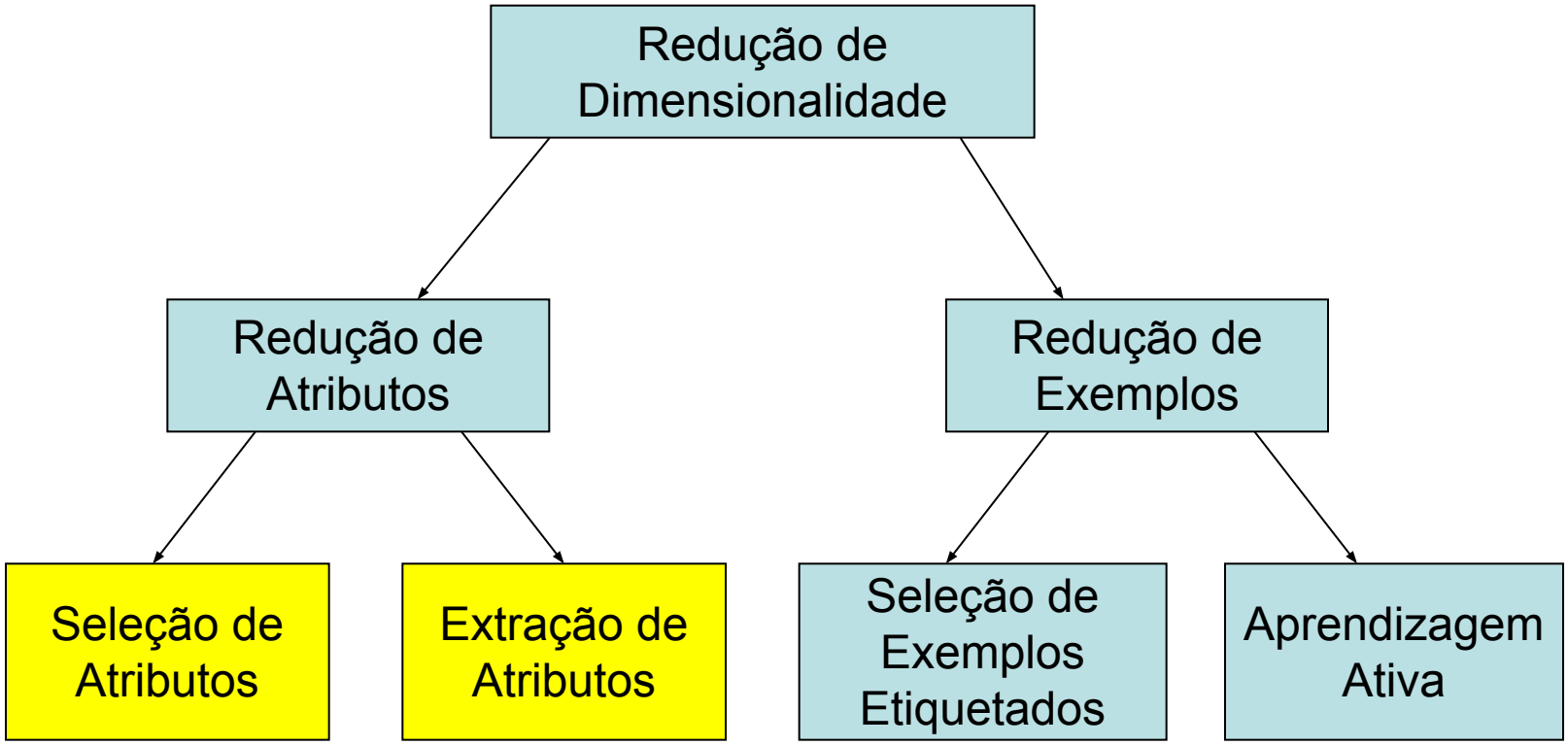
- Abordagens
  - Seleção X Extração de Atributos
- Seleção de atributos:
  - Escolha de um sub-conjunto de atributos relevantes dentre os atributos disponíveis
    - E.g., Filtros e Wrappers
- Extração de atributos:
  - Criação de novos atributos a partir da combinação dos atributos existentes
    - E.g., PCA

Redução de Exemplos

- Conjuntos de treinamento também podem apresentar exemplos irrelevantes e redundantes

- Abordagens:
  - Seleção de exemplos etiquetados
  - Seleção de exemplos não-etiquetados
    - *Aprendizagem Ativa*

Redução de Dimensionalidade



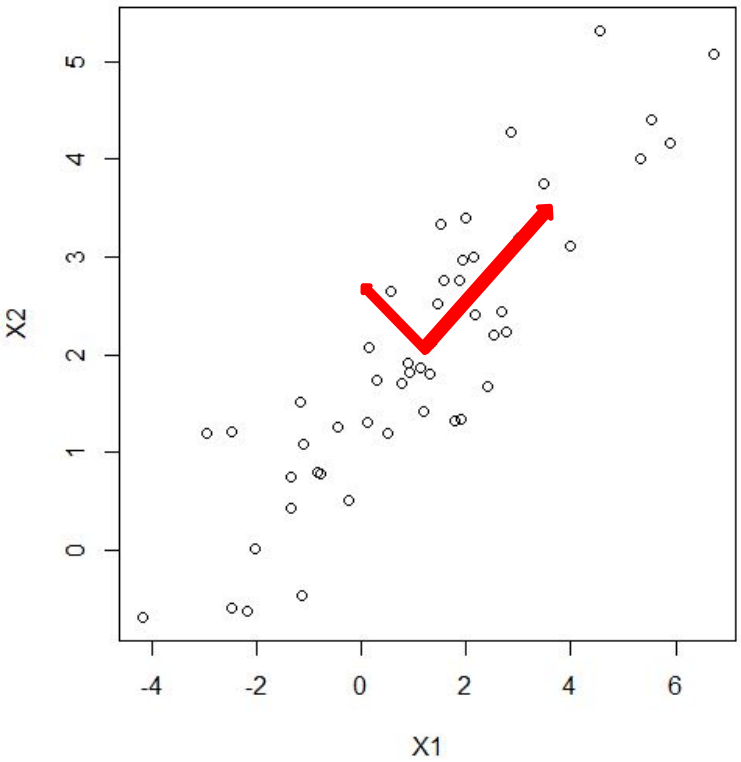
Extração de Atributos

- Criação de atributos relevantes através da combinação dos atributos originais
- Muito usado para análise exploratória de dados
- Diversas técnicas
  - PCA, ICA, SVD, FPCA,…

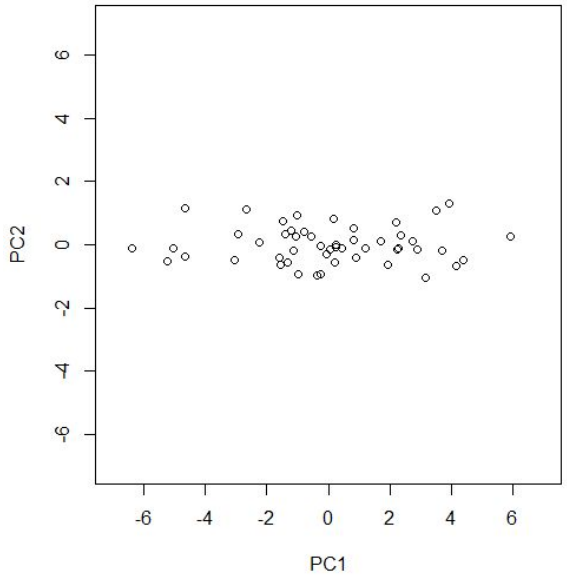
Principal Component Analysis

- Cada nova dimensão é uma combinação linear das variáveis originais dos dados
- Cada nova dimensão é chamada de componente principal
- Componentes principais são ordenados de forma a explicar a variância dos dados

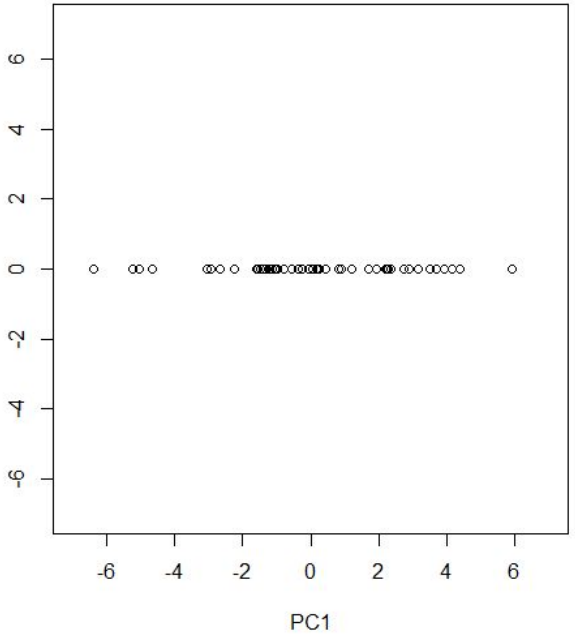
Dados Originais



Transformação PCA



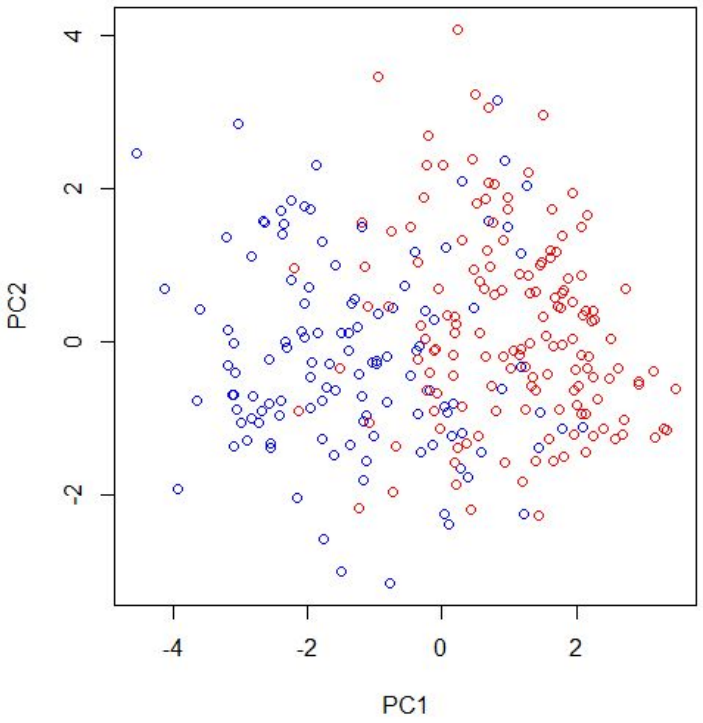
Redução de Dimensionalidade

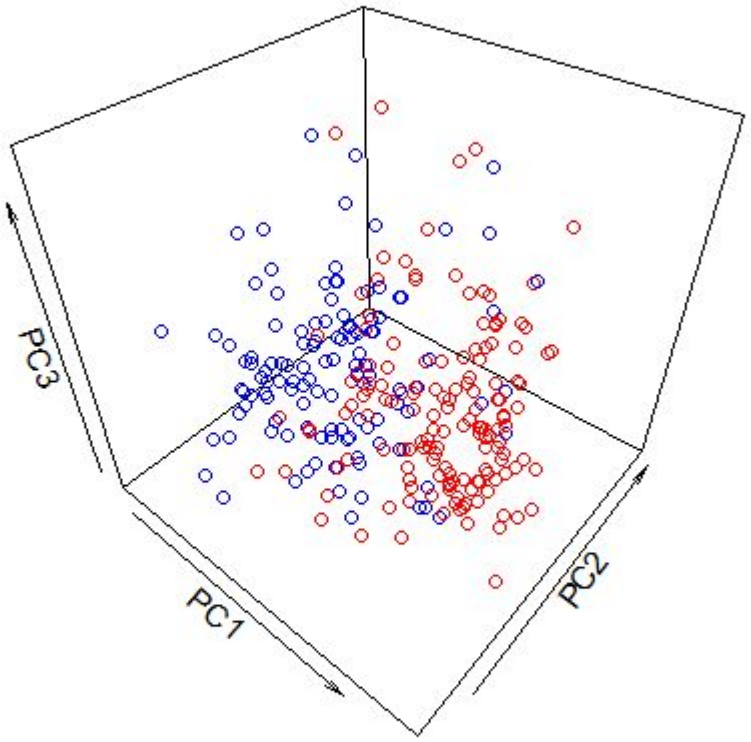


Heart Dataset

- 13 atributos

Standard deviations:
[1] 1.7450842 1.2679385 1.1140628 1.0773529 1.0061711 0.9532100 0.9087842
[8] 0.8656492 0.8231727 0.7590373 0.6734324 0.6346909 0.5821222

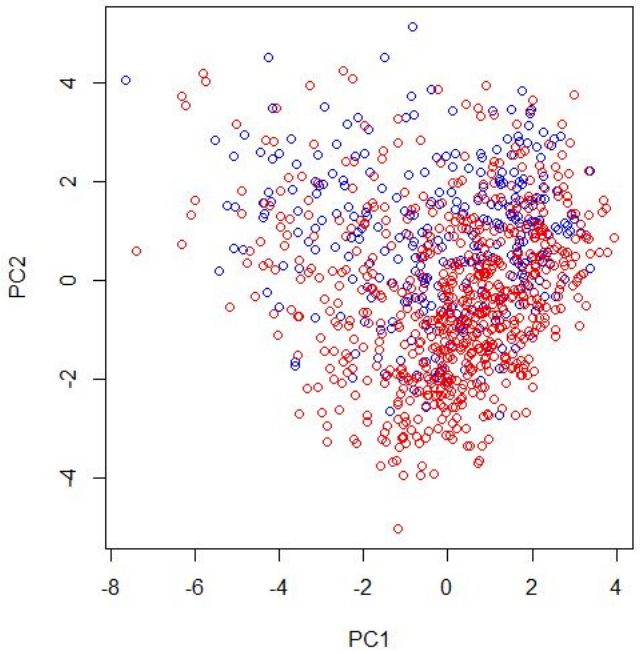


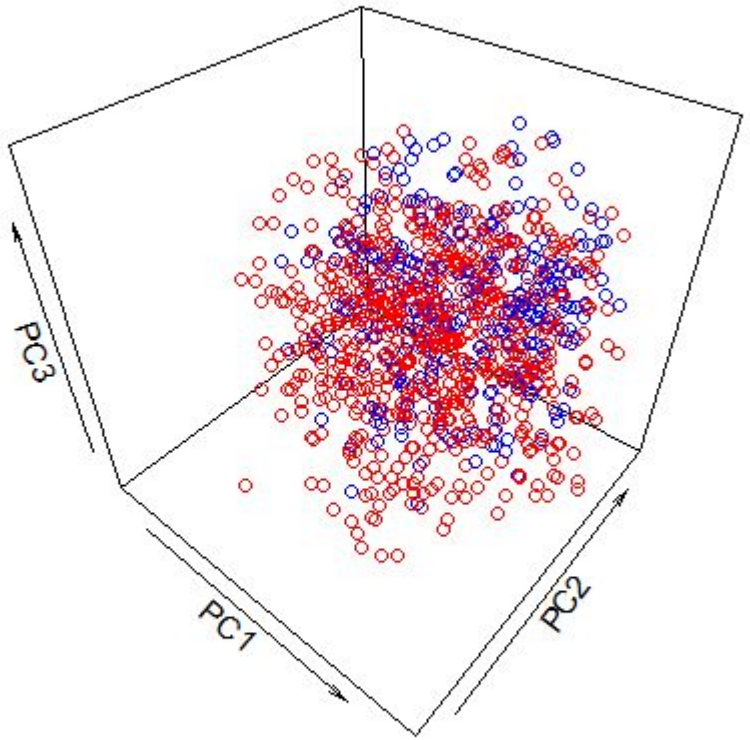


German Dataset

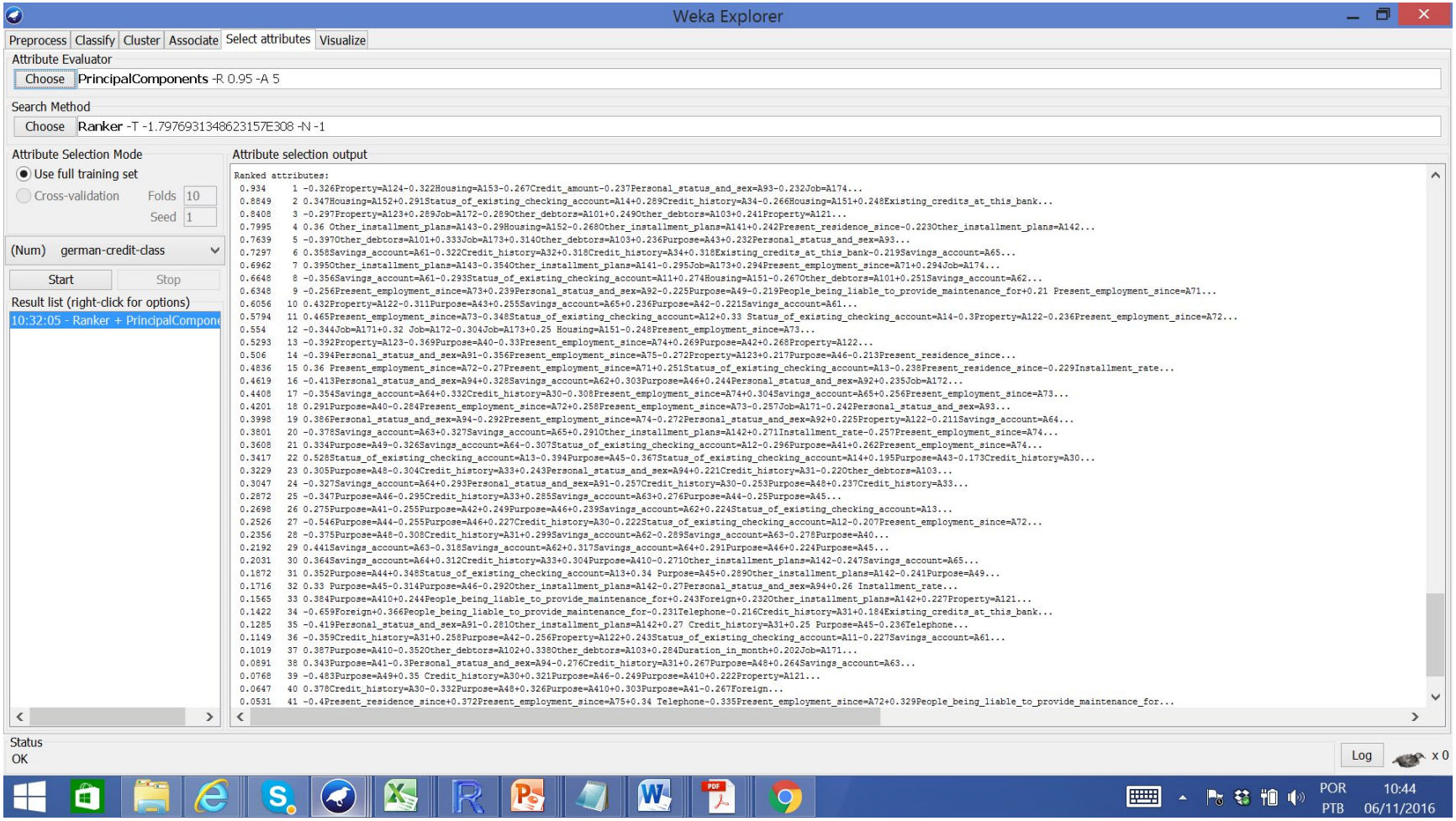
- 59 atributos

Standard deviations:
[1] 1.973460 1.701656 1.613873 1.560912 1.447674 1.421420 1.406148 1.361314
[9] 1.330906 1.310890 ...





PCA no WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
PrincipalComponents -R 0.95 -A 5
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Num) german-credit-class
Start
Stop
Result list (right-click for options)
10:32:05 - Ranker + PrincipalCompone
Attribute selection output
Ranked attributes:
0.934 1 -0.326Property=A124-0.322Housing=A153-0.267Credit_amount-0.237Personal_status_and_sex=A93-0.232Job=A174...
0.8849 2 0.347Housing=A152+0.291Status_of_existing_checking_account=A14+0.289Credit_history=A34-0.266Housing=A151+0.248Existing_credits_at_this_bank...
0.8408 3 -0.297Property=A123+0.289Job=A172-0.289Other_debtors=A101+0.249Other_debtors=A103+0.241Property=A121...
0.7995 4 0.36 Other_installment_plans=A143-0.29Housing=A152-0.268Other_installment_plans=A141+0.242Present_residence_since-0.223Other_installment_plans=A142...
0.7639 5 -0.397Other_debtors=A101+0.333Job=A173+0.314Other_debtors=A103+0.236Purpose=A43+0.232Personal_status_and_sex=A93...
0.7297 6 0.358Savings_account=A61-0.322Credit_history=A32+0.318Credit_history=A34+0.318Existing_credits_at_this_bank-0.219Savings_account=A65...
0.6962 7 0.395Other_installment_plans=A143-0.354Other_installment_plans=A141-0.295Job=A173+0.294Present_employment_since=A71+0.294Job=A174...
0.6648 8 -0.356Savings_account=A61-0.293Status_of_existing_checking_account=A11+0.274Housing=A151-0.267Other_debtors=A101+0.251Savings_account=A62...
0.6348 9 -0.256Present_employment_since=A73+0.239Personal_status_and_sex=A92-0.225Purpose=A49-0.219People_being_liable_to_provide_maintenance_for+0.21 Present_employment_since=A71...
0.6056 10 0.432Property=A122-0.311Purpose=A43+0.255Savings_account=A65+0.236Purpose=A42-0.221Savings_account=A61...
0.5794 11 0.465Present_employment_since=A73-0.348Status_of_existing_checking_account=A12+0.33 Status_of_existing_checking_account=A14-0.3Property=A122-0.236Present_employment_since=A72...
0.554 12 -0.344Job=A171+0.32 Job=A172-0.304Job=A173+0.25 Housing=A151-0.248Present_employment_since=A73...
0.5293 13 -0.392Property=A123-0.369Purpose=A40-0.33Present_employment_since=A74+0.269Purpose=A42+0.268Property=A122...
0.506 14 -0.394Personal_status_and_sex=A91-0.356Present_employment_since=A75-0.272Property=A123+0.217Purpose=A46-0.213Present_residence_since...
0.4836 15 0.36 Present_employment_since=A72-0.27Present_employment_since=A71+0.251Status_of_existing_checking_account=A13-0.238Present_residence_since-0.229Installment_rate...
0.4619 16 -0.413Personal_status_and_sex=A94+0.328Savings_account=A62+0.303Purpose=A46+0.244Personal_status_and_sex=A92+0.235Job=A172...
0.4408 17 -0.354Savings_account=A64+0.332Credit_history=A30-0.308Present_employment_since=A74+0.304Savings_account=A65+0.256Present_employment_since=A73...
0.4201 18 0.291Purpose=A40-0.284Present_employment_since=A72+0.258Present_employment_since=A73-0.257Job=A171-0.242Personal_status_and_sex=A93...
0.3998 19 0.386Personal_status_and_sex=A94-0.292Present_employment_since=A74-0.272Personal_status_and_sex=A92+0.225Property=A122-0.211Savings_account=A64...
0.3801 20 -0.378Savings_account=A63+0.327Savings_account=A65+0.291Other_installment_plans=A142+0.271Installment_rate-0.257Present_employment_since=A74...
0.3608 21 0.334Purpose=A49-0.326Savings_account=A64-0.307Status_of_existing_checking_account=A12-0.296Purpose=A41+0.262Present_employment_since=A74...
0.3417 22 0.528Status_of_existing_checking_account=A13-0.394Purpose=A45-0.367Status_of_existing_checking_account=A14+0.195Purpose=A43-0.173Credit_history=A30...
0.3229 23 0.305Purpose=A48-0.304Credit_history=A33+0.243Personal_status_and_sex=A94+0.221Credit_history=A31-0.22Other_debtors=A103...
0.3047 24 -0.327Savings_account=A64+0.293Personal_status_and_sex=A91-0.257Credit_history=A30-0.253Purpose=A48+0.237Credit_history=A33...
0.2872 25 -0.347Purpose=A46-0.295Credit_history=A33+0.285Savings_account=A63+0.276Purpose=A44-0.25Purpose=A45...
0.2698 26 0.275Purpose=A41-0.255Purpose=A42+0.249Purpose=A46+0.239Savings_account=A62+0.224Status_of_existing_checking_account=A13...
0.2526 27 -0.546Purpose=A44-0.255Purpose=A46+0.227Credit_history=A30-0.222Status_of_existing_checking_account=A12-0.207Present_employment_since=A72...
0.2356 28 -0.375Purpose=A48-0.308Credit_history=A31+0.299Savings_account=A62-0.289Savings_account=A63-0.278Purpose=A40...
0.2192 29 0.441Savings_account=A63-0.318Savings_account=A62+0.317Savings_account=A64+0.291Purpose=A46+0.224Purpose=A45...
0.2031 30 0.364Savings_account=A64+0.312Credit_history=A33+0.304Purpose=A410-0.271Other_installment_plans=A142-0.247Savings_account=A65...
0.1872 31 0.352Purpose=A44+0.348Status_of_existing_checking_account=A13+0.34 Purpose=A45+0.289Other_installment_plans=A142-0.241Purpose=A49...
0.1716 32 0.33 Purpose=A45-0.314Purpose=A46-0.292Other_installment_plans=A142-0.27Personal_status_and_sex=A94+0.26 Installment_rate...
0.1565 33 0.384Purpose=A410+0.244People_being_liable_to_provide_maintenance_for+0.243Foreign+0.232Other_installment_plans=A142+0.227Property=A121...
0.1422 34 -0.659Foreign+0.366People_being_liable_to_provide_maintenance_for-0.231Telephone-0.216Credit_history=A31+0.184Existing_credits_at_this_bank...
0.1285 35 -0.419Personal_status_and_sex=A91-0.281Other_installment_plans=A142+0.27 Credit_history=A31+0.25 Purpose=A45-0.236Telephone...
0.1149 36 -0.359Credit_history=A31+0.258Purpose=A42-0.256Property=A122+0.243Status_of_existing_checking_account=A11-0.227Savings_account=A61...
0.1019 37 0.387Purpose=A410-0.352Other_debtors=A102+0.338Other_debtors=A103+0.284Duration_in_month+0.202Job=A171...
0.0891 38 0.343Purpose=A41-0.3Personal_status_and_sex=A94-0.276Credit_history=A31+0.267Purpose=A48+0.264Savings_account=A63...
0.0768 39 -0.483Purpose=A49+0.35 Credit_history=A30+0.321Purpose=A46-0.249Purpose=A410+0.222Property=A121...
0.0647 40 0.378Credit_history=A30-0.332Purpose=A48+0.326Purpose=A410+0.303Purpose=A41-0.267Foreign...
0.0531 41 -0.4Present_residence_since+0.372Present_employment_since=A75+0.34 Telephone-0.335Present_employment_since=A72+0.329People_being_liable_to_provide_maintenance_for...
Status
OK
Log
x 0
POR
PTB
10:44
06/11/2016

Seleção de Atributos

- Filtros
  - Atributos são ordenados com base em *métricas* de relevância e redundância
  - Retorna os atributos mais bem orderados

- Wrappers
  - O desempenho do algoritmo é avaliado para diferentes sub-conjuntos de atributos;
  - O melhor sub-conjunto encontrado é retornado;

Seleção de Atributos – Filtros

- Descartam atributos irrelevantes antes do processo de aprendizado
  - I.e., independente do algoritmo
- Caracteristicas gerais dos dados são levadas em consideração para selecionar atributos
  - I.e. dados avaliados com estatísticas descritivas
- Diferentes métricas podem ser usadas para definir relevância de atributos
  - K atributos mais relevantes são retornados

Seleção de Atributos – Filtros

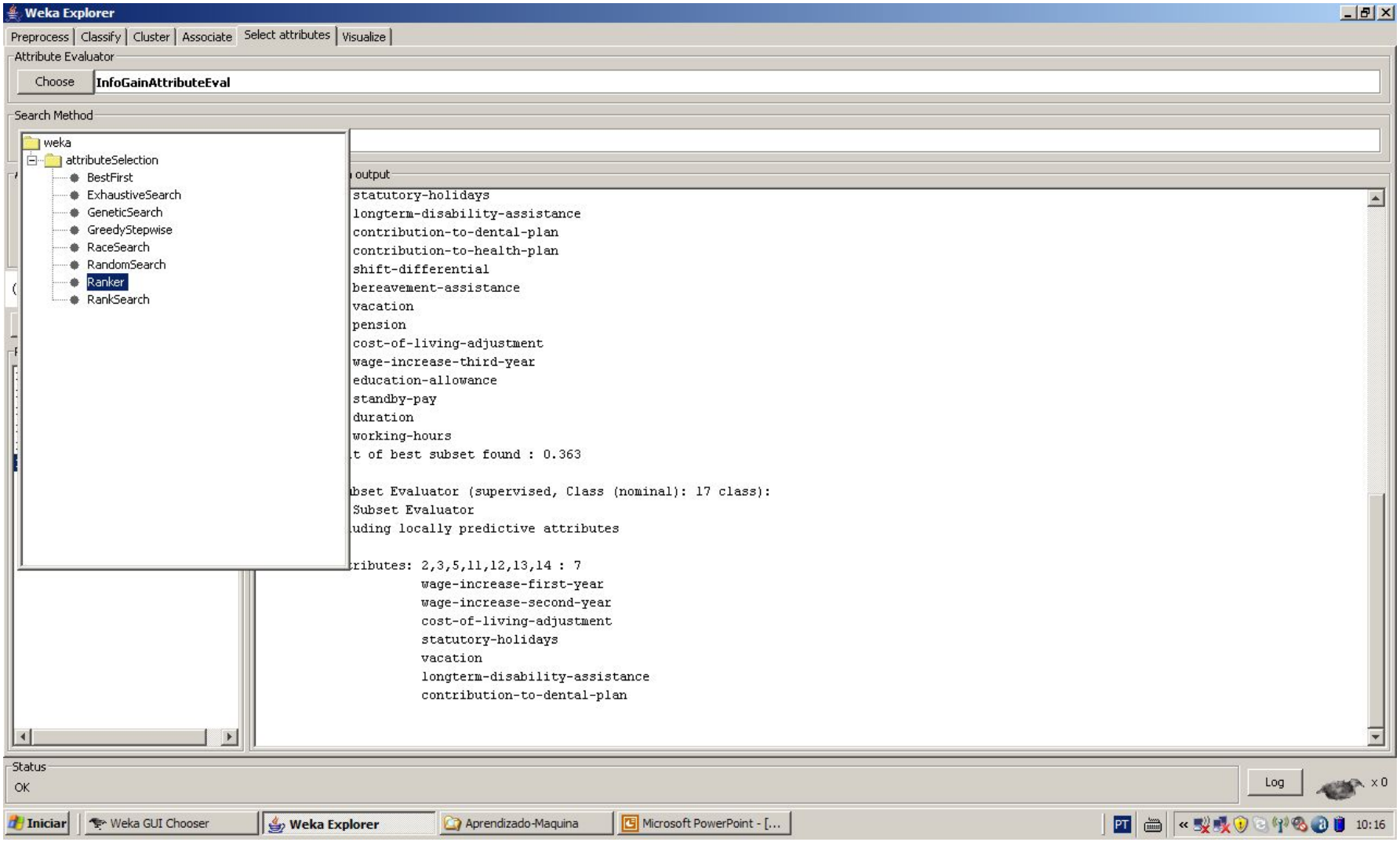
- Exemplos de métricas:
  - InfoGain, GainRatio,… e outras métricas usadas para seleção de atributos em árvores de decisão
  - *Correlação* entre o atributo avaliado e o atributo classe

Seleção de Atributos – Filtros

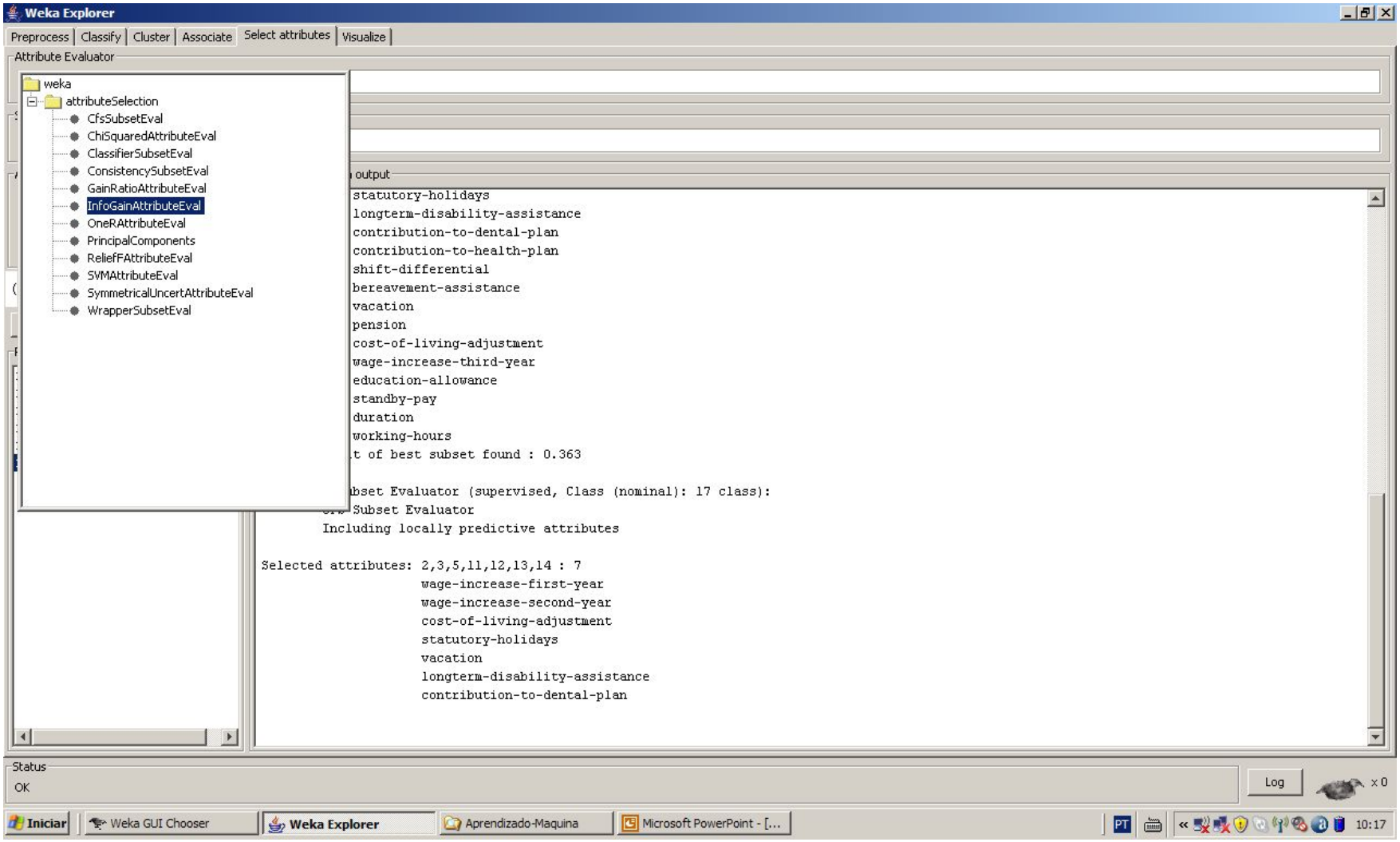
- Em geral, filtros de atributos são leves computacionalmente

- Porém apresenta dificuldades:
  - Nem sempre é fácil definir quantos atributos descartar;
    - Na prática, isso é definido por tentativa-e-erro
  - Não leva em consideração o algoritmo sendo utilizado;

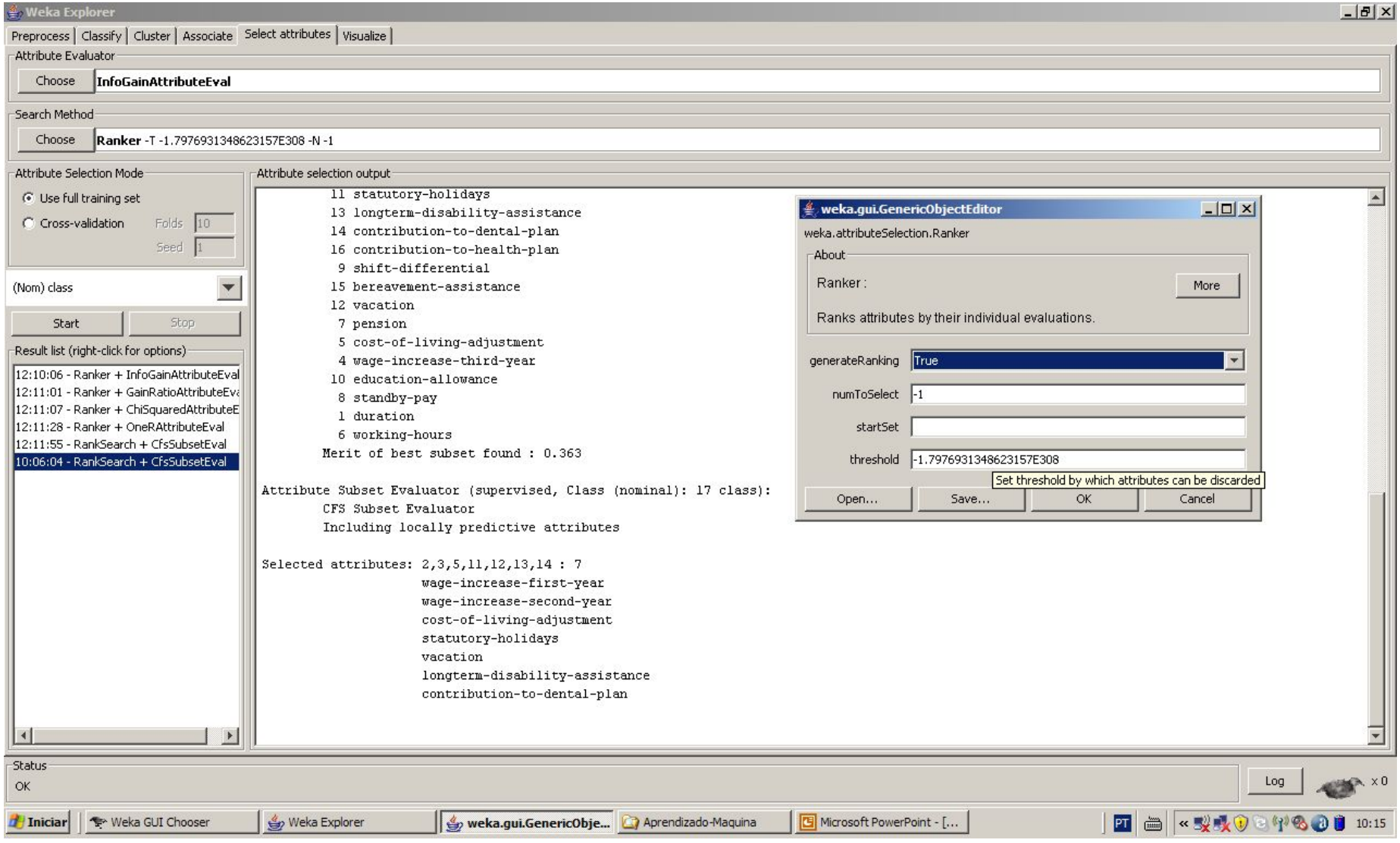
Seleção de Atributos – Filtros (WEKA)



Seleção de Atributos – Filtros (WEKA)



Seleção de Atributos – Filtros (WEKA)



Seleção de Atributos – Wrappers

- Wrappers realizam, de fato, uma *busca* no espaço de sub-conjuntos de atributos
  - Busca exaustiva não é realizada na prática

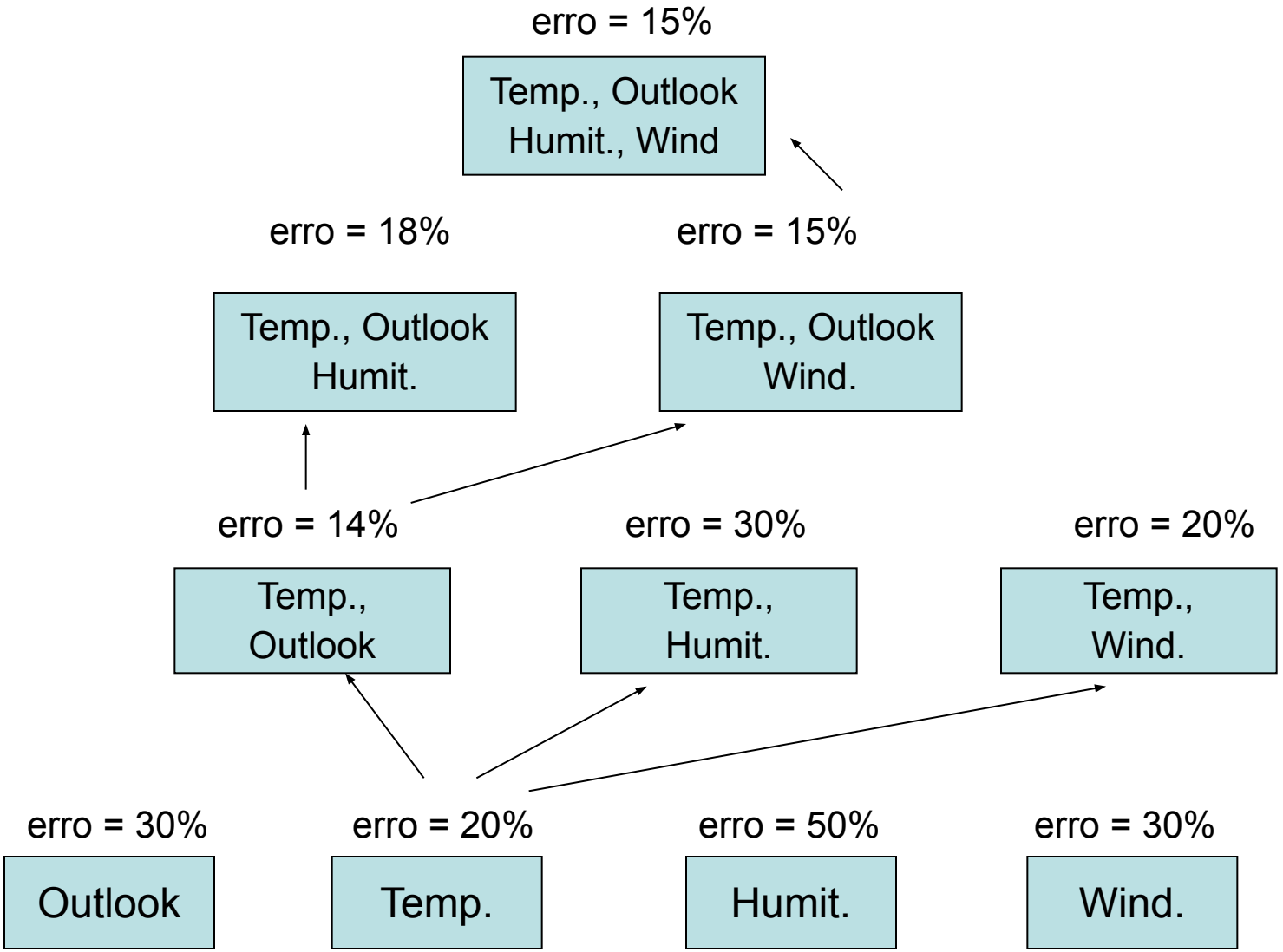
- São vantajosos em relação aos filtros por considerarem o algoritmo em questão

Seleção de Atributos – Wrappers

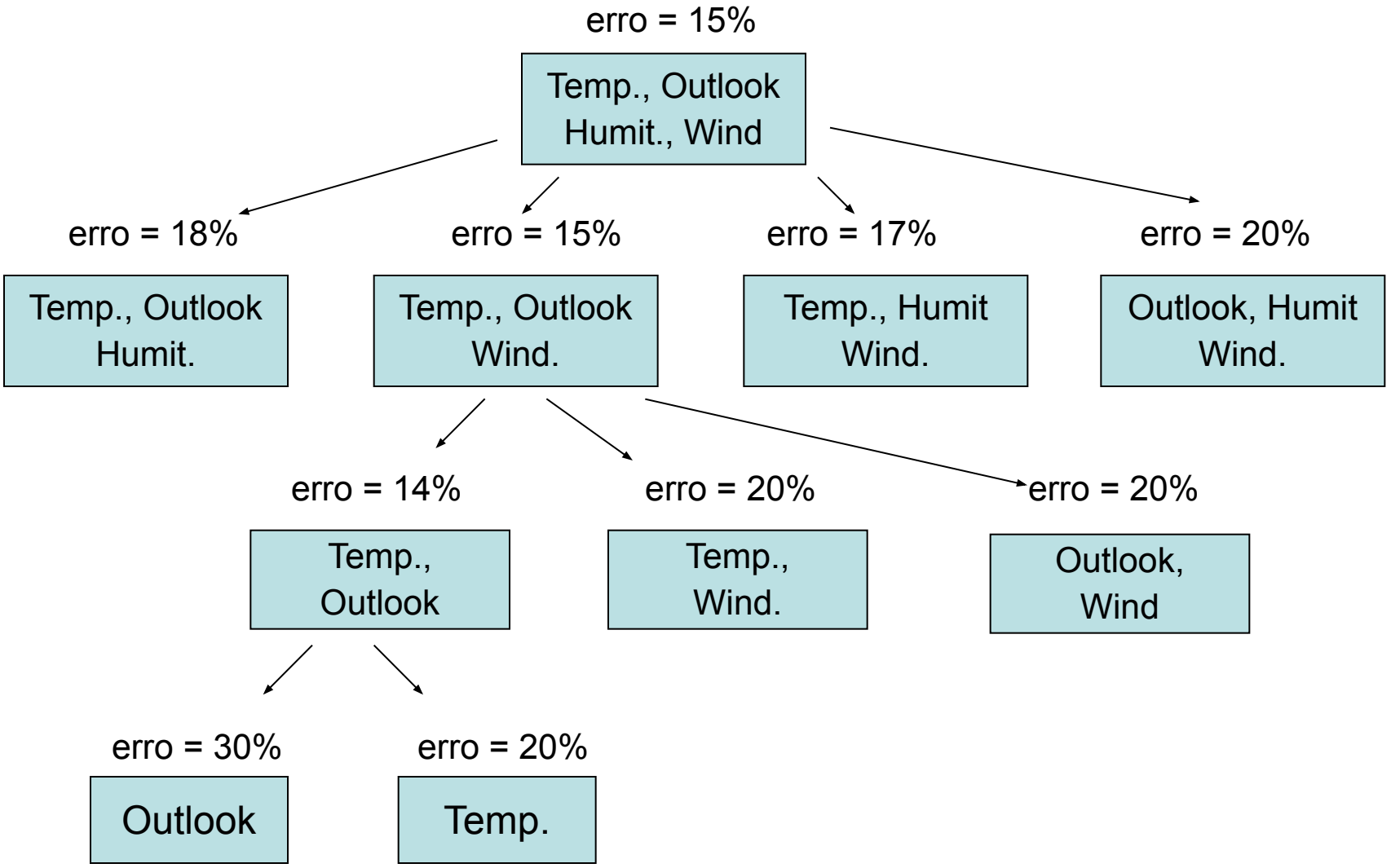
- Forward-Selection
  - Busca se inicia a partir de sub-conjuntos com um atributo
  - Atributos são adicionados progressivamente

- Backward-Elimination
  - Inicia com o sub-conjunto de todos os atributos;
  - Atributos são removidos progressivamente

Wrappers – Forward-Selection



Wrappers – Backward-Elimination



Seleção de Atributos – Wrappers

- Backward-Elimination
  - Em geral, produz melhores resultados em termos de precisão;
  - Porém é mais pesado computacionalmente

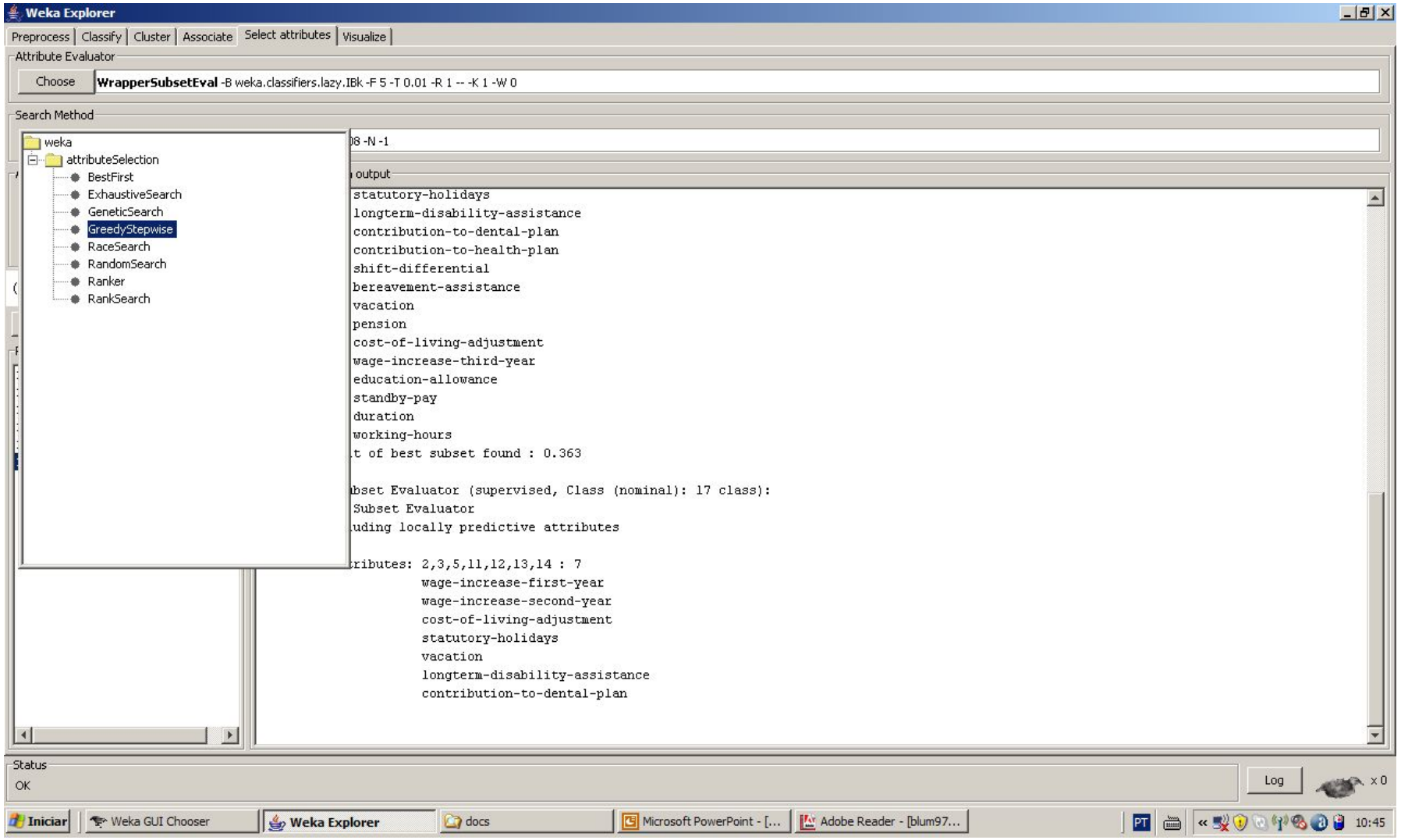
- Forward-Selecion
  - Tende a produzir sub-conjuntos com menos atributos;
  - É capaz de eliminar melhor atributos redundantes

Seleção de Atributos – Wrappers

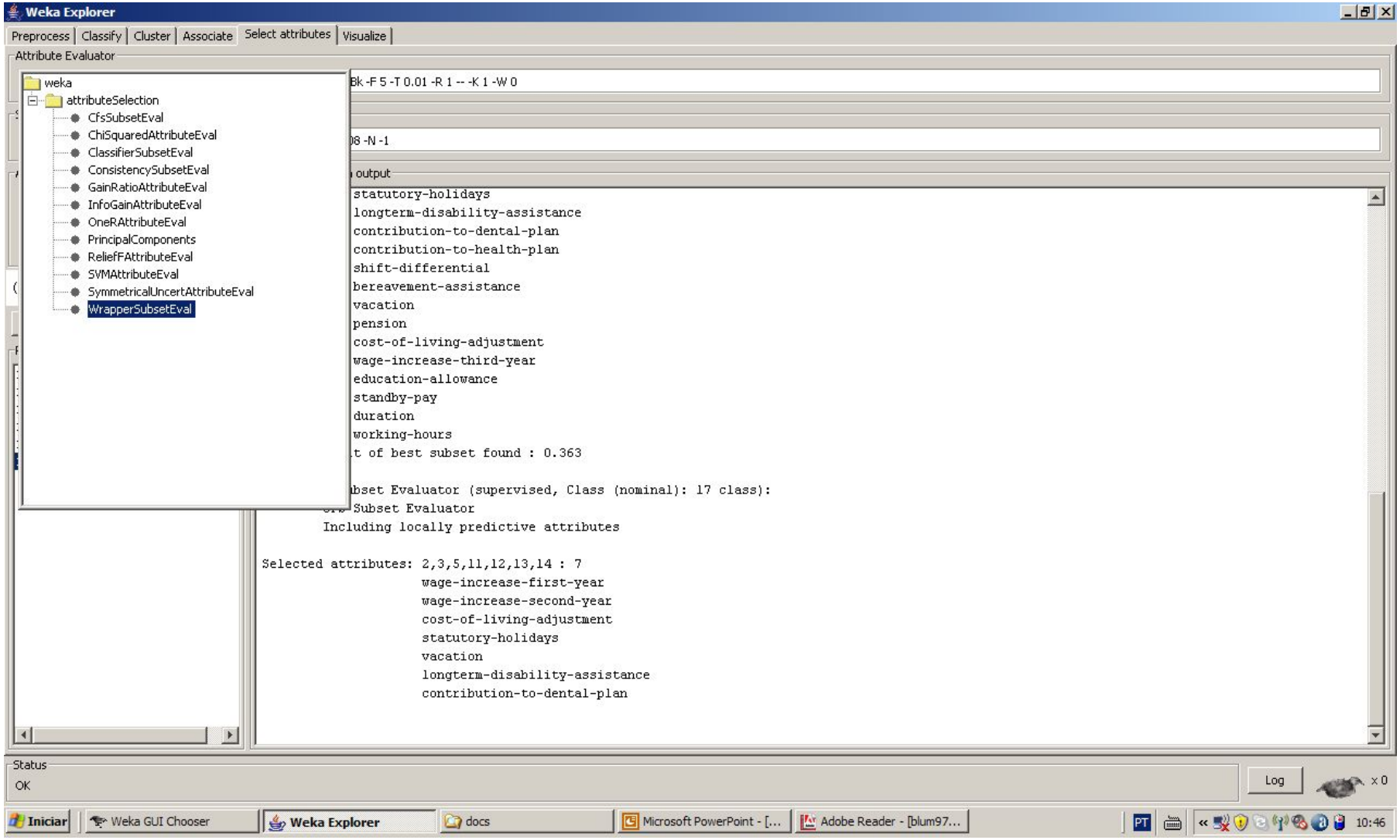
- Tanto Forward-Selection como Backward-Elimination caem em *mínimos locais*
  - Ambos os algoritmos realizam *greedy search*

- Diversos algoritmos de busca e otimização podem ser aplicados
  - E.g., Algoritmos Genéticos

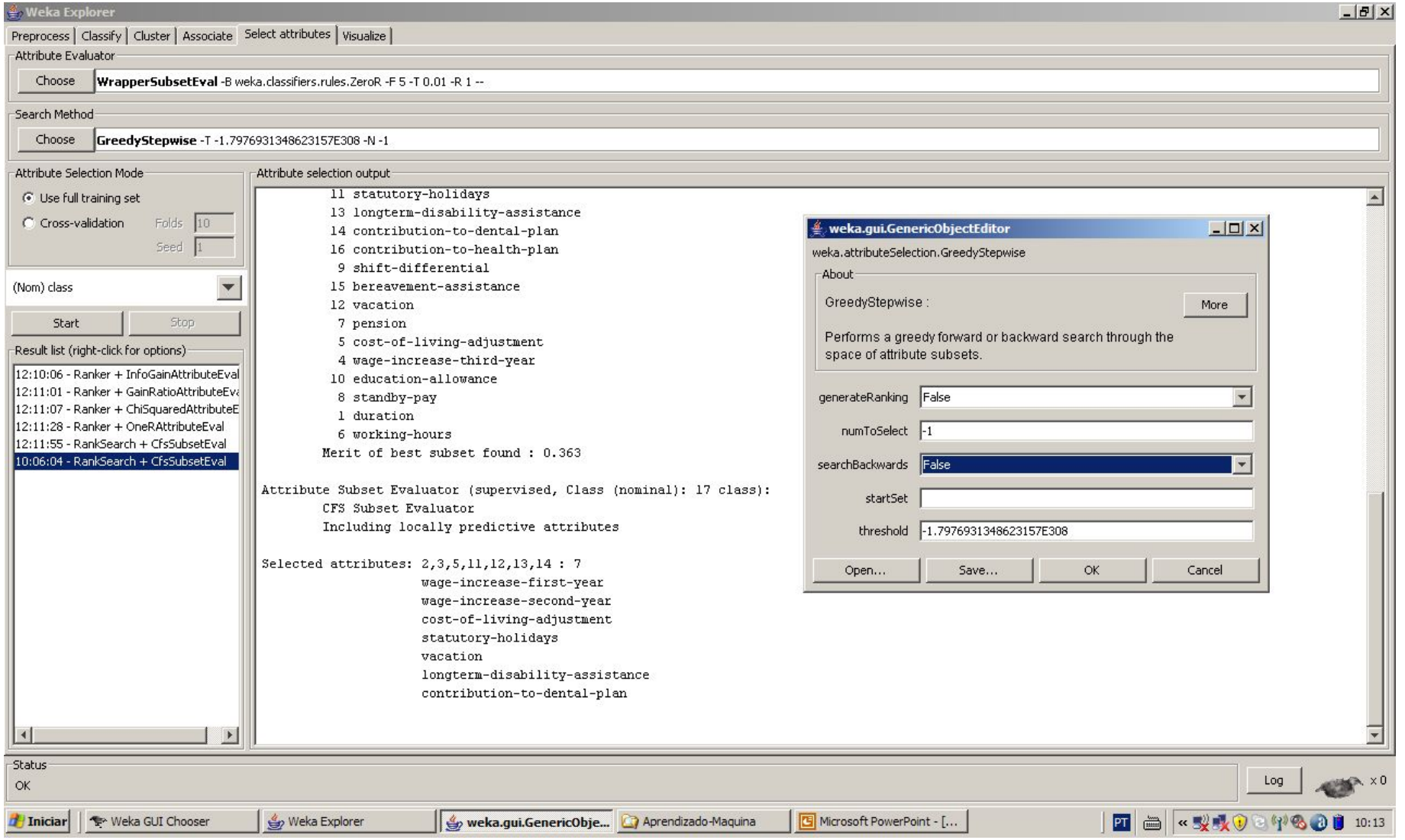
Seleção Atributos – Wrappers (WEKA)



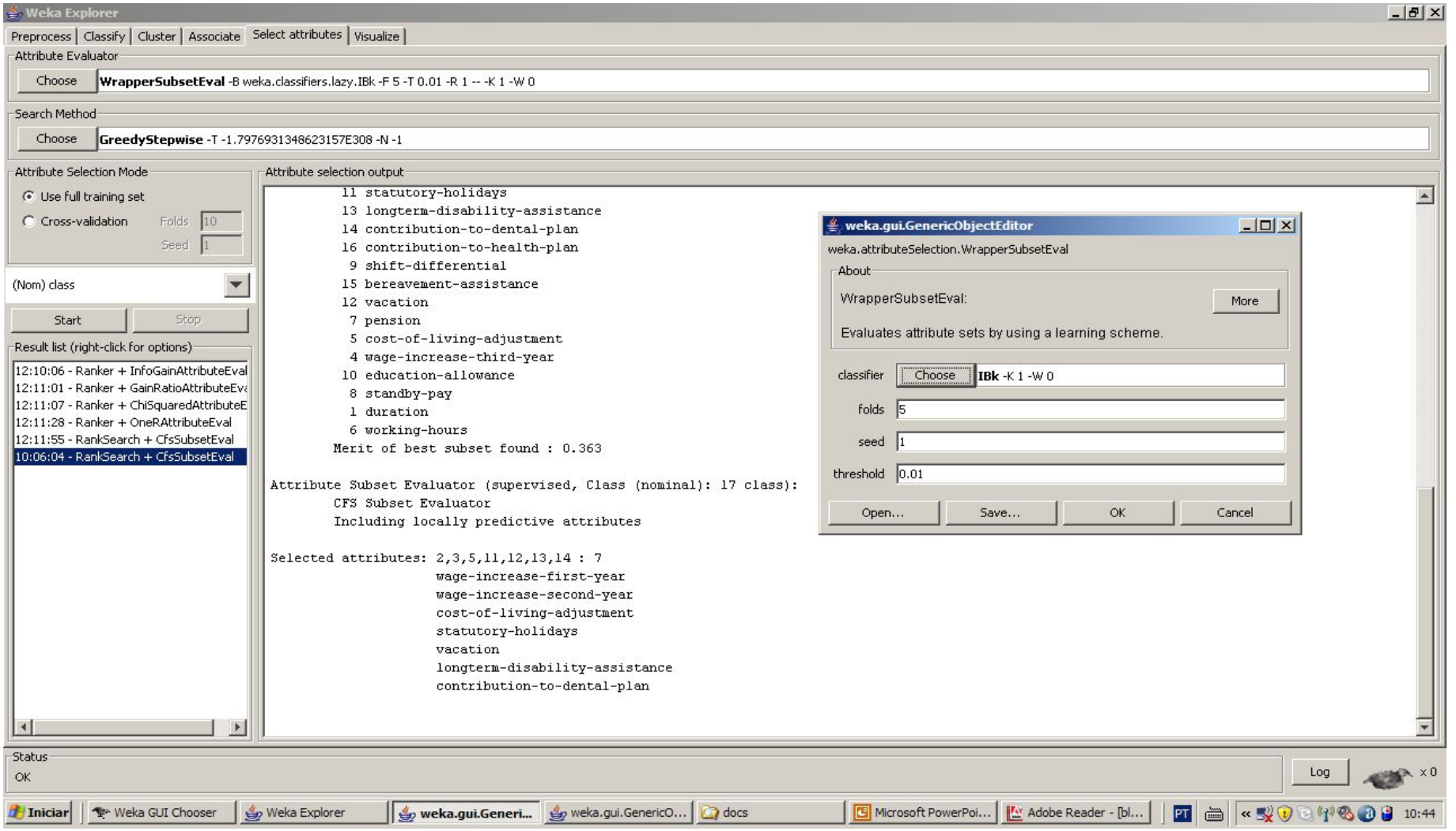
Seleção Atributos – Wrappers (WEKA)



Seleção Atributos – Wrappers (WEKA)



Seleção Atributos – Wrappers (WEKA)



Seleção de Atributos

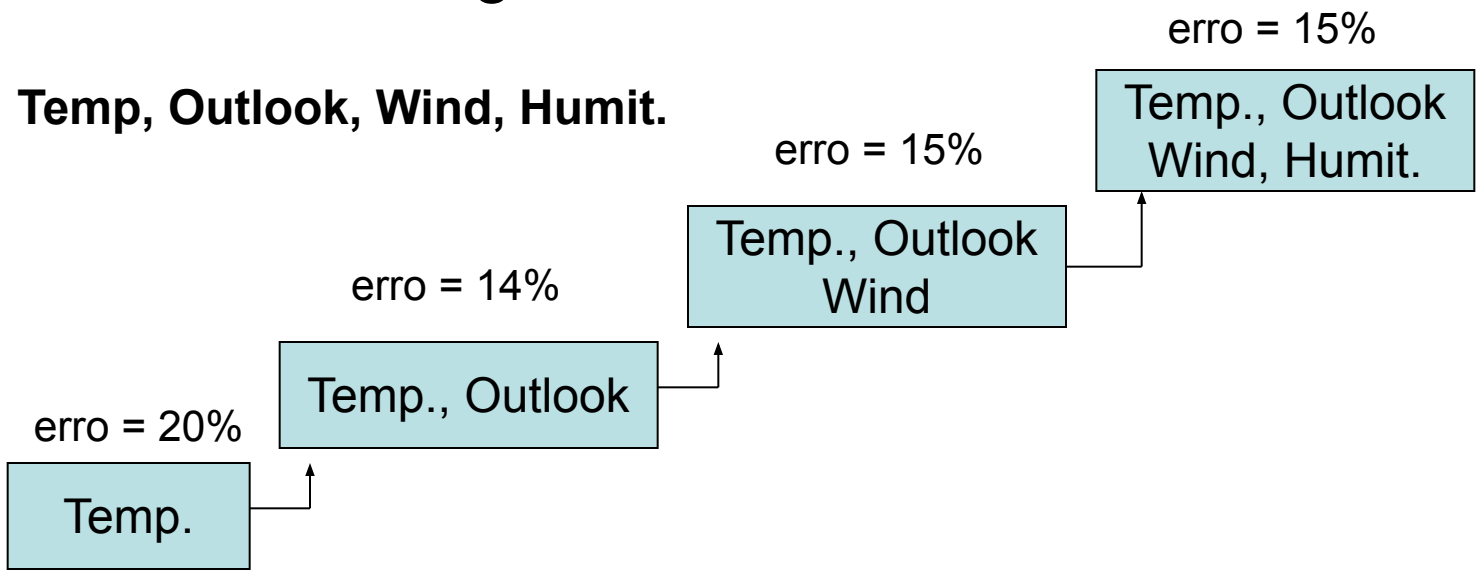
- Busca em espaço de atributos é mais pesado computacionalmente que as técnicas de filtros

- Porém, podemos minimizar esse problema:
  - Utilizando algoritmos de busca mais eficientes (e.g., *RankSearch*);
  - Utilizando algoritmos mais leves para avaliação dos sub-conjuntos;
  - Combinando com filtros (e.g., usar uma métrica de filtro como medida de avaliação)

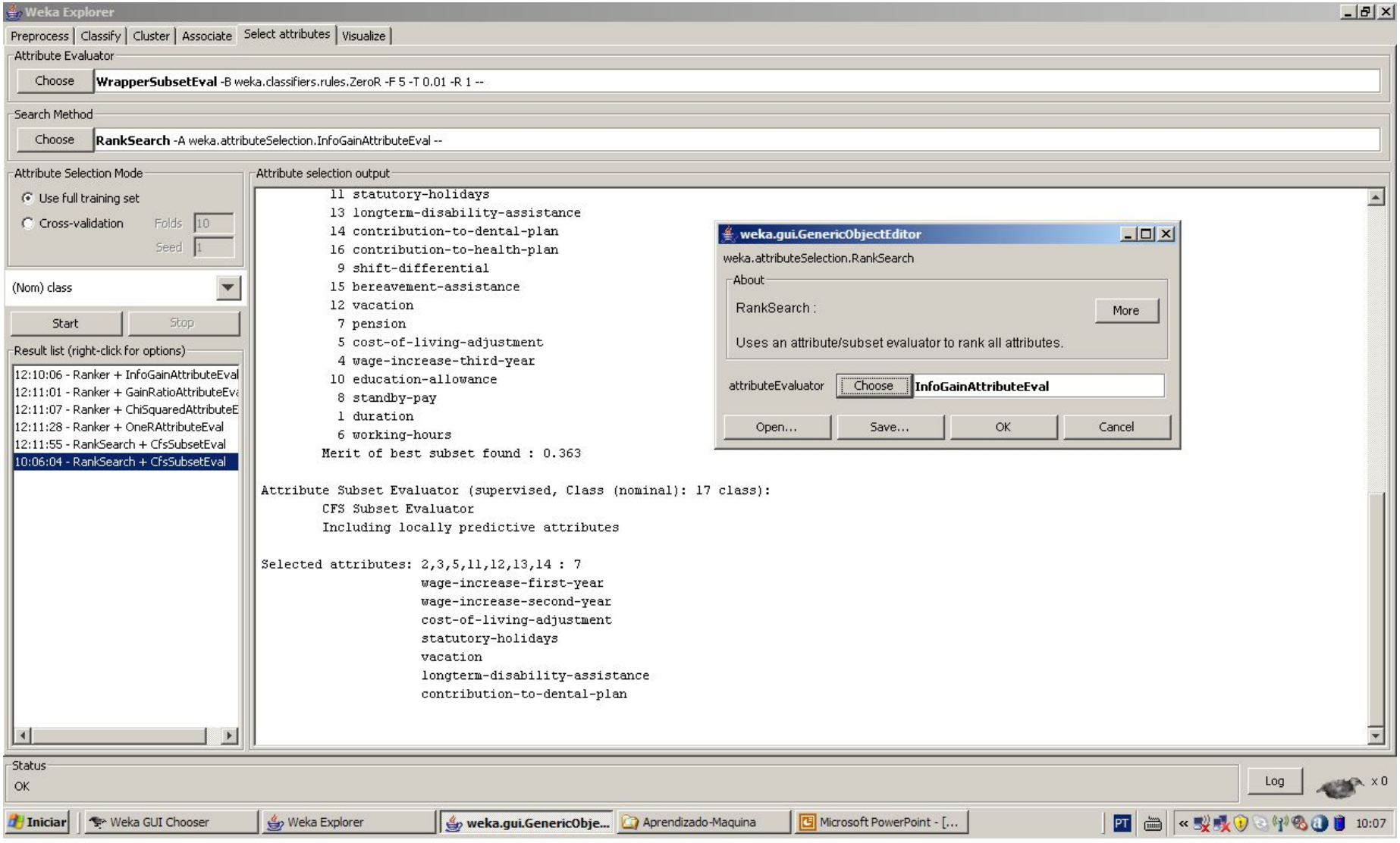
Seleção de Atributos

- RankSearch
  - Ordenar atributos com uma métrica de filtro
    - E.g., InfoGain
  - Avaliar sub-conjuntos, inserindo gradualmente os atributos na ordem gerada

Ranking: Temp, Outlook, Wind, Humit.



Seleção de Atributos – RankSearch (WEKA)



Referências

- A. Blum, P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence*, 1997.
- R. Kohavi, GH John, Wrappers for feature subset selection, *Artificial Intelligence*, 1997.
- I. Guyon, A. Elisseefi, L. Kaelbling, An Introduction to variable and feature selection, *Journal of Machine Learning Research*, 2003.