Seleção de Atributos via Informação Estatística

Celio Henrique Nogueira Larcher Junior

Laboratório Nacional de Computação Científica

Petrópolis, 2017

Agenda

Agenda

Introdução

- Técnicas para seleção de atributos é a denominação de um grupo de técnicas que buscam extrair subconjuntos de atributos significativos da base de dados
- A seleção de atributos é uma operação importante no contexto de aprendizado de máquina:
  - Permite a simplificação de modelos
  - Diminui o tempo de treinamento
  - Ameniza a "maldição da dimensionalidade"
  - Reduz o problema de *overfitting*

Definição

- Dado uma base de dados formada pelo conjunto de atributos S, tem-se como objetivo encontrar um subconjunto $K \subseteq S$ suficientemente representativo:
  - $|K|$ deve ser o menor possível
  - Os atributos de K devem ser capazes de conter informação semelhante ao visto em S
- Em particular para classificação/regressão espera-se que K tenha tanta informação sobre a saída C quanto S

Agenda

Critério da máxima dependência

- A aplicação de informação mútua a um conjunto de atributos pode ser um bom indicativo da adequabilidade dos mesmos:

$$I(K; C) = I(\{F_1, F_2, \ldots, F_{|K|}\}; C)$$

- Desta forma, tem-se um indicativo do quanto o conjunto de atributos tem de informação em comum com os valores de saída (C).

Critério da máxima dependência - Problemas

- Porém o número de atributos não é penalizado, não sendo desencorajada a escolha de atributos redundantes.
- Além disto, calcular a informação mútua de um conjunto de atributos pode ser custoso.
- Necessária a aplicação da regra da cadeia, ou outro mecanismo que o valha.

$$I(K;C) = \sum_{i=1}^{|K|} I(F_i; C|F_{i-1}, F_{i-2}, \ldots, F_1) = I(K)+I(C)-I(K,C)$$

- Complexidade cresce proporcionalmente ao produto das dimensão dos espaços de atributos (produto cartesiano)

mRMR - min-Redundancy Max-Relevance

- Uma opção menos custosa é o método mRMR:

$$I(F_j; C) - \frac{1}{|K|} \sum_{F_i \in K} I(F_j; F_i)$$

- Pondera a informação ganha pela adição de um novo atributo com a redundância para os atributos já selecionados no conjunto K.
- Ainda possui complexidade considerável (proporcional ao número de atributos já selecionado).

Ganho de informação

- Como uma alternativa de custo consideravelmente mais baixo tem-se a métrica do ganho de informação:

$$G_i = I(C) - I(C|F_i) = I(C; F_i)$$

- O ganho de informação é dado pela diferença entre a entropia para a classe de saída e a entropia ao se utilizar o atributo F_i.
- Não há qualquer consideração em relação aos atributos já selecionados.

Informação normalizada

- Ainda outra alternativa semelhante é a a informação normalizada:

$$R_i = \frac{I(C) - I(C|F_i)}{I(F_i)} = \frac{I(C;F_i)}{I(F_i)}$$

- Considera atributos de menor informação própria como mais relevantes para seleção.

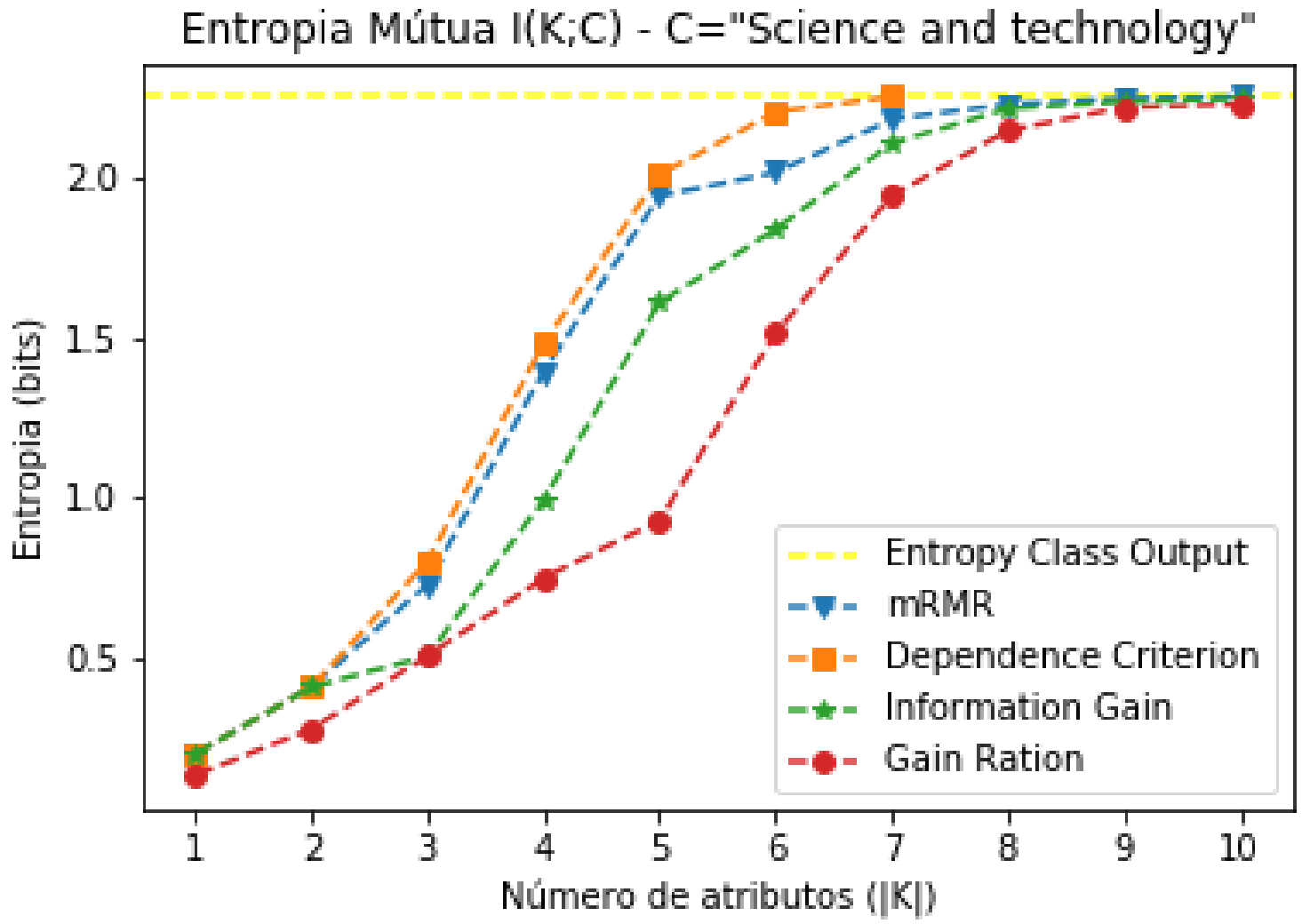
Exploração do espaço de busca

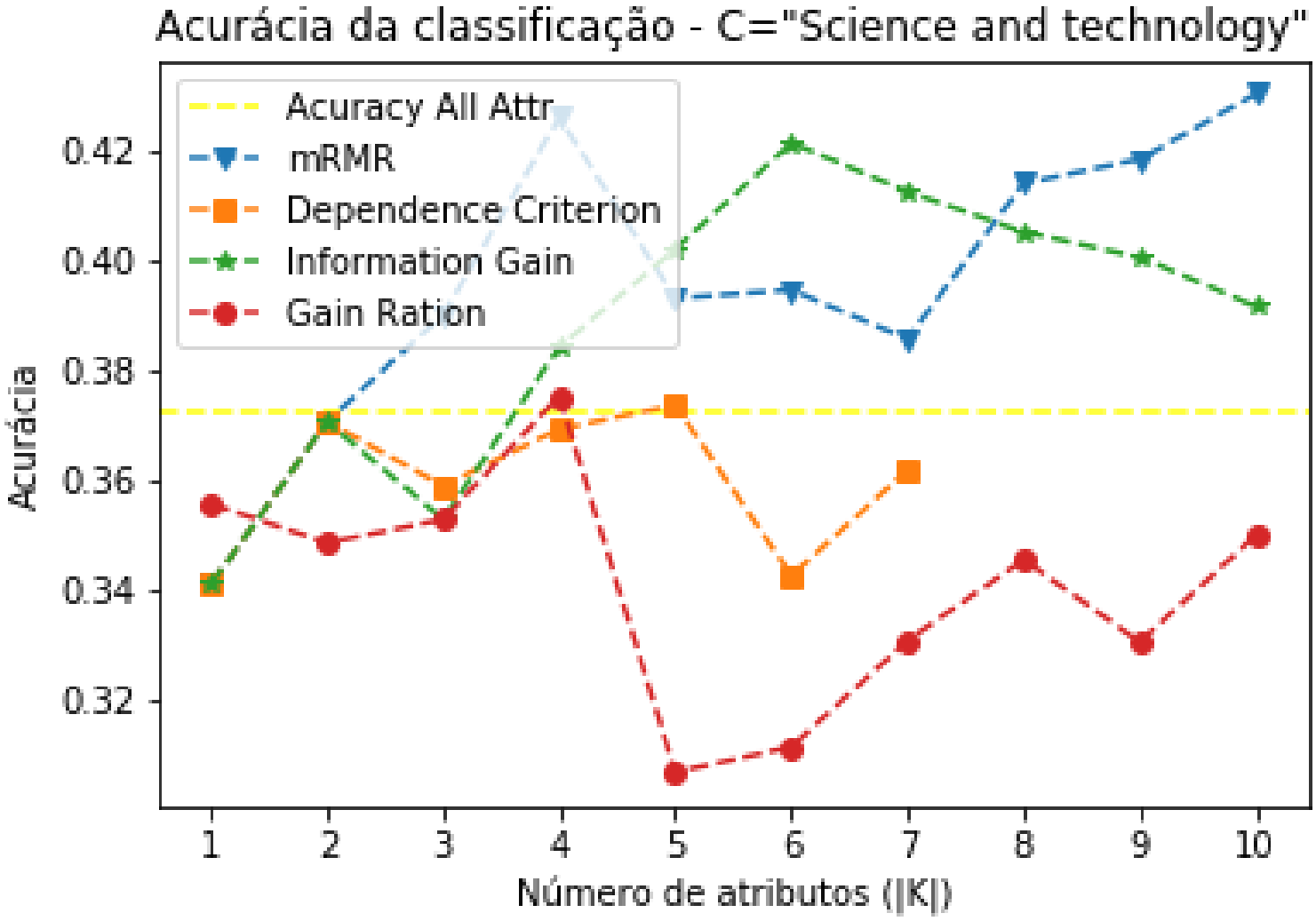
- Outro ponto a ser definido além das métricas para seleção são os métodos de exploração.
- Espaço total para exploração é exponencial ($2^{|S|}$)
- Opções para exploração:
  - Exploração completa (caro)
  - Procedimento guloso (sub-ótimo)
  - Procedimentos adaptativos (metaheurísticas)

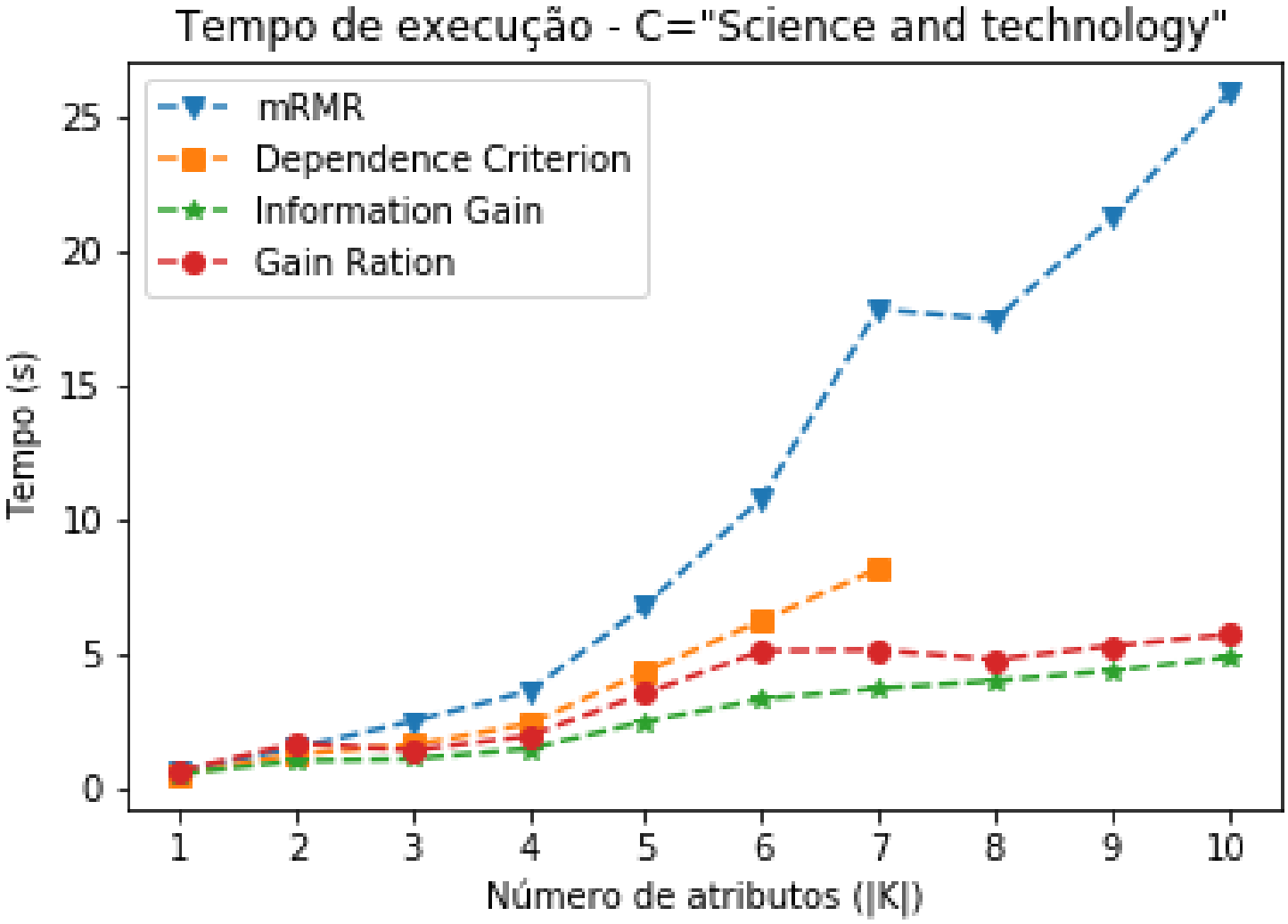
Agenda

Configuração

- Para realização dos experimentos foi selecionado a base de dados *Young People Survey* do repositório *Kaggle*
- A base consiste de 150 atributos e 1010 registros
- Foi implementado um procedimento de seleção guloso para seleção dos atributos
- Para avaliação do processo de classificação utilizou-se o algoritmo SVM

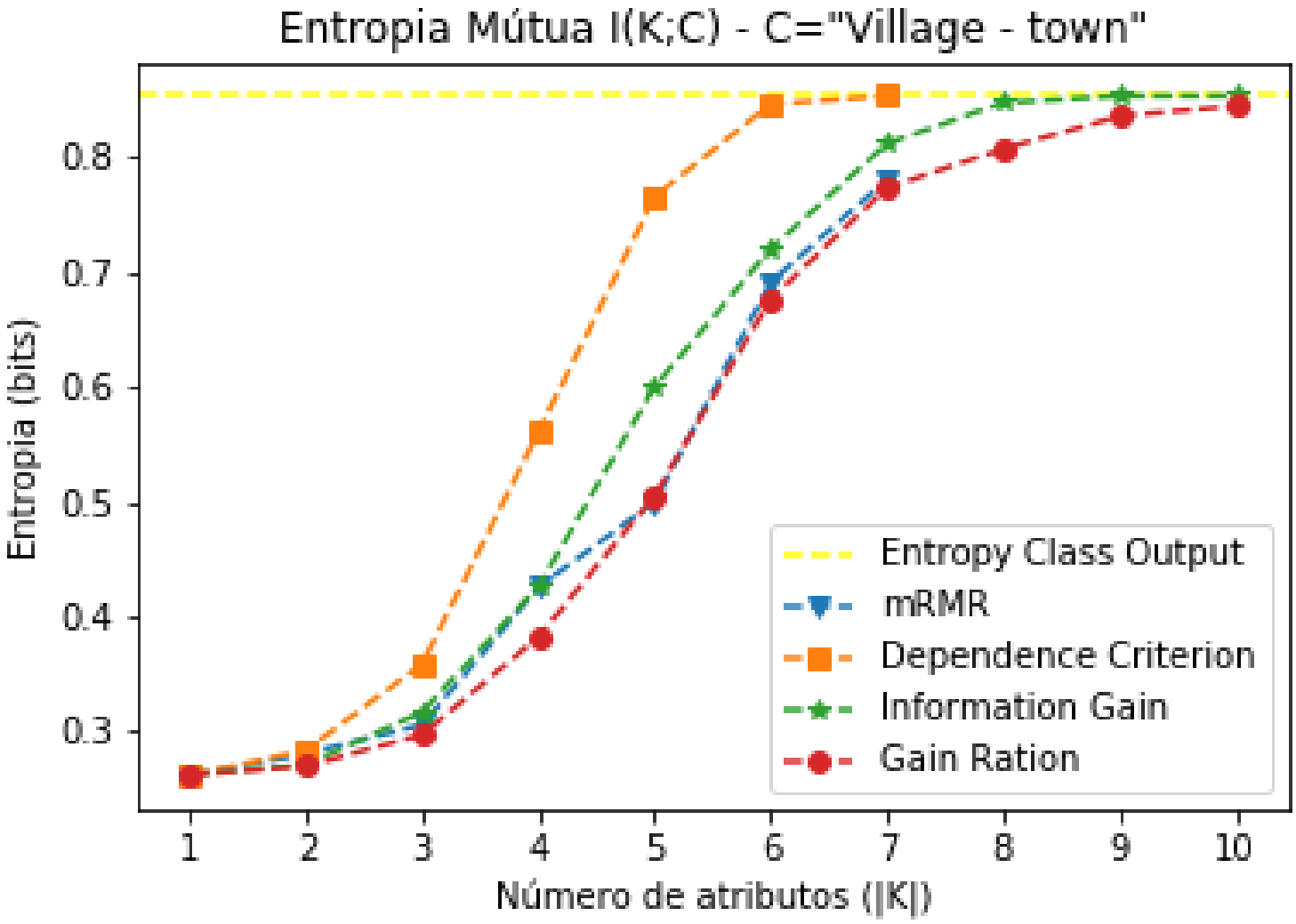

Entropia Mútua I(K;C) - C="Science and technology"
Entropia (bits)
2.0
1.5
1.0
0.5
1
2
3
4
5
6
7
8
9
10
Número de atributos (|K|)
Entropy Class Output
mRMR
Dependence Criterion
Information Gain
Gain Ration

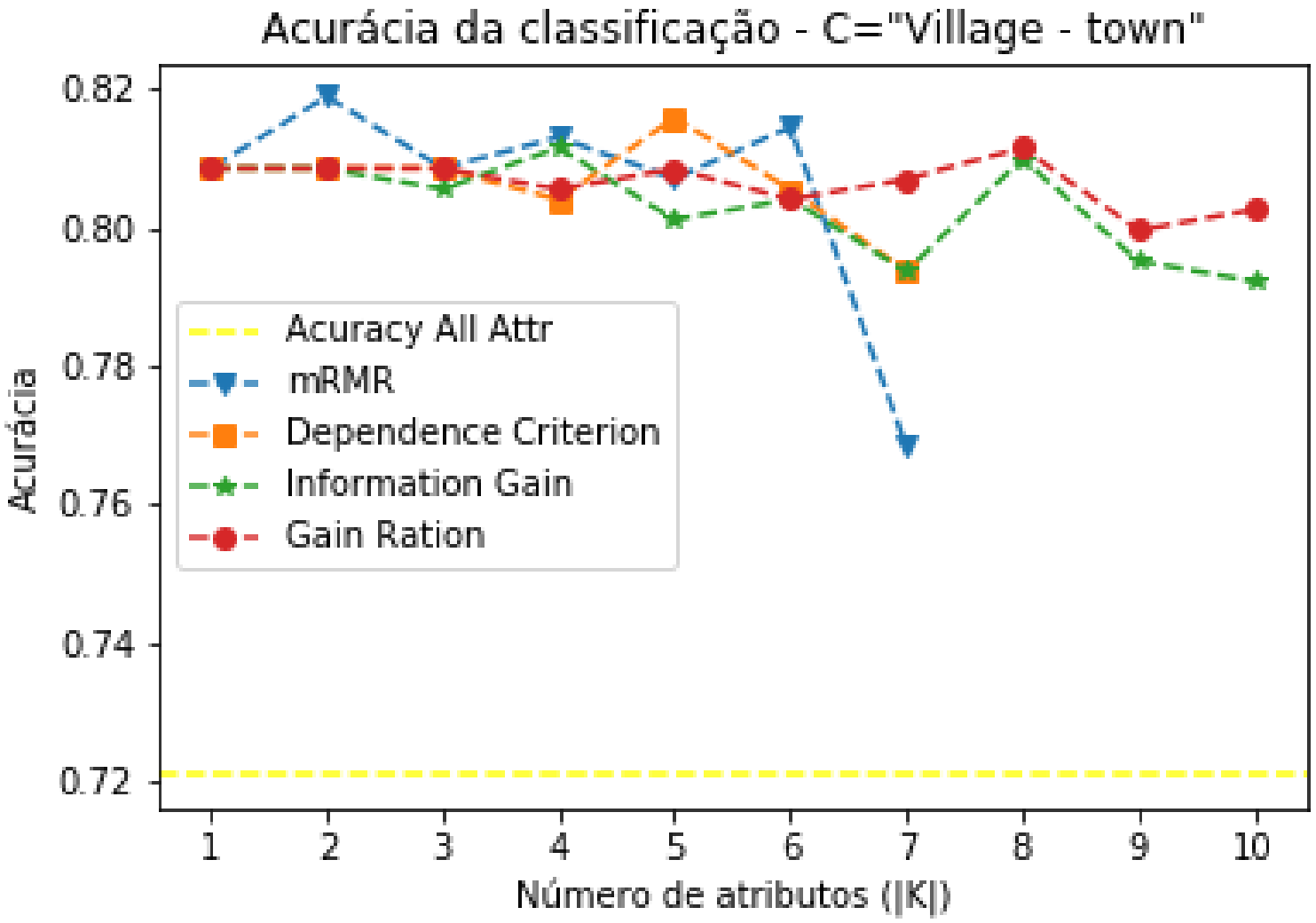

Acurácia da classificação - C="Science and technology"
Acuracy All Attr
mRMR
Dependence Criterion
Information Gain
Gain Ration
Acurácia
0.42
0.40
0.38
0.36
0.34
0.32
1
2
3
4
5
6
7
8
9
10
Número de atributos (|K|)

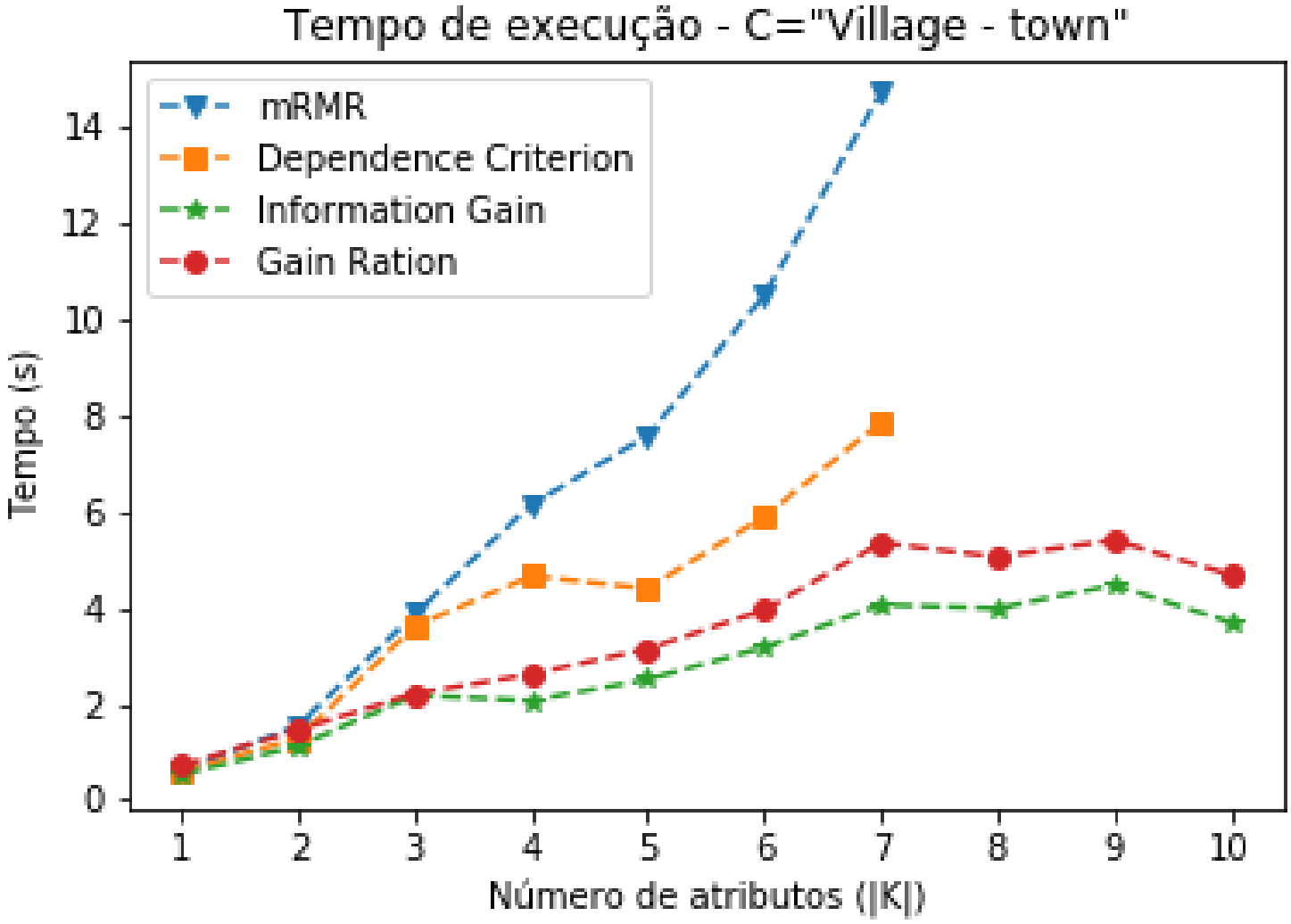

Tempo de execução - C="Science and technology"
mRMR
Dependence Criterion
Information Gain
Gain Ration
Tempo (s)
0
5
10
15
20
25
1
2
3
4
5
6
7
8
9
10
Número de atributos (|K|)

Atributos selecionados

- Mrmr: [PC, Physics, Documentary, Cars, Branded clothing, Gender, Sci-fi, Adrenaline sports, Daily events, Funniness]
- Dependence Criterion: [PC, Physics, Cars, Horror, Branded clothing, Spending on healthy eating, Pets]
- Information Gain: [PC, Physics, Gender, Cars, Documentary, Height, Sci-fi, Spending on gadgets, Life struggles, Action]
- Gain Ration: [Gender, Physics, PC, Height, Weight, Cars, Documentary, Sci-fi, Spending on gadgets, Western]


Entropia Mútua I(K;C) - C="Village - town"
Entropia (bits)
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
4
5
6
7
8
9
10
Número de atributos (|K|)
Entropy Class Output
mRMR
Dependence Criterion
Information Gain
Gain Ration


Acurácia da classificação - C="Village - town"
0.82
0.80
0.78
0.76
0.74
0.72
Acurácia
Acuracy All Attr
mRMR
Dependence Criterion
Information Gain
Gain Ration
1
2
3
4
5
6
7
8
9
10
Número de atributos (|K|)


Tempo de execução - C="Village - town"
mRMR
Dependence Criterion
Information Gain
Gain Ration
Tempo (s)
0
2
4
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
Número de atributos (|K|)

Atributos selecionados

- Mrmr: [House - block of flats, Friends versus money, Punctuality, Geography, Only child, Punk, Height]
- Dependence Criterion: [House - block of flats, Punk, Passive sport, War, Rats, Biology, Public speaking]
- Information Gain: [House - block of flats, Religion, Gardening, Friends versus money, Number of siblings, Storm, Countryside - outdoors, Punk, Final judgement, God]
- Gain Ration: [House - block of flats, Movies, Gardening, Religion, Number of siblings, Storm, Friends versus money, Punctuality, Countryside - outdoors, Personality]

Agenda

Sobre o experimento

- A informação mútua não é diretamente proporcional à acurácia obtida (dependente da capacidade de generalização/representação da técnica de aprendizado)
- Espaço amostral pequeno: 1010 registros X $\approx 5^{150}$ possibilidades)
- Tempo de execução obtido não é absoluto (carece de otimizações)
- Método de seleção guloso não garante a melhor escolha possível para cada métrica

Sobre as técnicas

- Mais atributos podem levar a um modelo mais complexo, mas de menor acurácia
- A possibilidade de redução para a dimensionalidade é substancial para ambas as tarefas de classificação analisadas
- A técnica máxima dependência possui uma convergência mais rápida em relação a entropia, porém com desempenho inferior para acurácia dos modelos gerados
- A técnica mRMR foi a que demonstrou um melhor balanço entre acurácia e entropia, porém foi a de maior tempo de execução
- As técnicas de Ganho de Informação e Informação Normalizada não apresentam comportamento constante para acurácia, mas Ganho de Informação tem maior crescimento da entropia

Agenda

Referências I


Bonev, B. I. (2010).
Feature Selection Based on Information Theory. *PHD Thesis.*


Duch, W., Biesiada, J., Winiarski, T., Grudziński, K., Grabczewski, K. (2003).
Feature Selection Based on Information Theory Filters. *Neural Networks and Soft Computing: Proceedings of the Sixth International Conference on Neural Networks and Soft Computing, Zakopane, Poland, June 11–15, 2002.*


Kaggle - Young People Survey.
https://www.kaggle.com/miroslavsabo/young-people-survey


Vergara, J. , Estévez, P. (2014).
A review of feature selection methods based on mutual information. *Neural Computing and Applications.*

Obrigado pela atenção!