Métodos para Seleção de Atributos em Mineração de Dados

Stanley Robson de M. Oliveira

Agenda

- **Seleção de atributos**:
  - Necessidade, motivação e objetivos.
- **Abordagens para seleção de atributos**:
  - Métodos não-Supervisionados;
  - Métodos Supervisionados;
  - Métodos estatísticos.
- **Aspectos relevantes**:
  - Estudo de caso;
  - Benchmark (comparação de métodos);
  - Limitações;
  - Desafios de pesquisa.

A seleção de variáveis é sempre necessária?

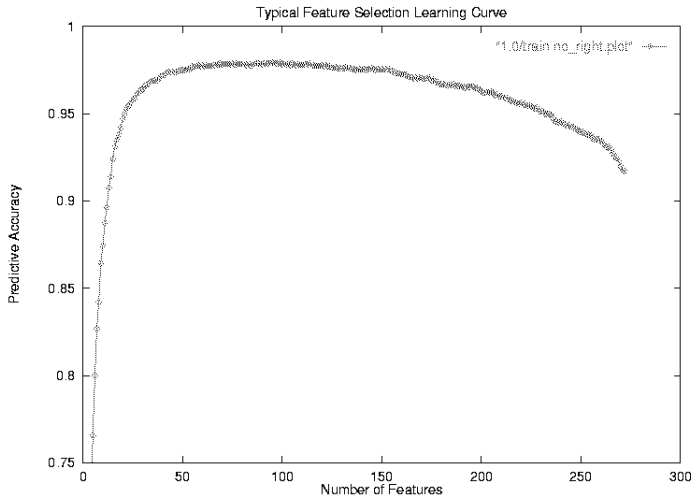
- Alguns **métodos** de aprendizado fazem **seleção de atributos** implicitamente.
- **Árvores de decisão**:
  - Usam **ganho de informação** para decidir quais os atributos serão considerados pela árvore gerada.
  - Alguns atributos podem não ser usados.
- **Redes Neurais**:
  - ***Backpropagation*** "aprende" fortes conexões para algumas entradas (**inputs**); e
  - Fracas conexões (**near-zero**) para outras entradas.

A seleção de variáveis é sempre necessária?

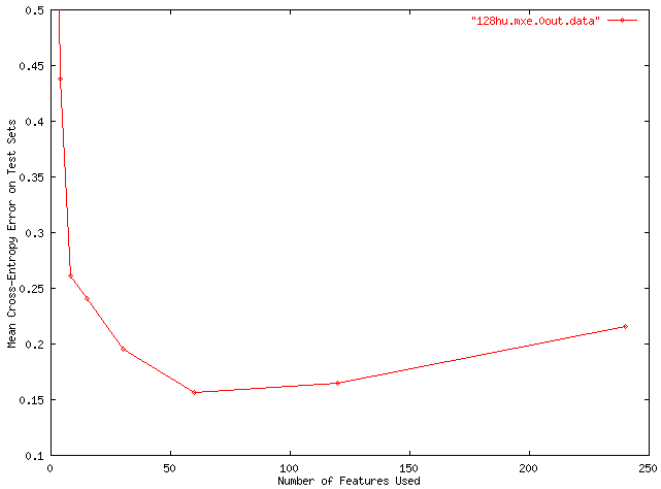
- **kNN, MBL**:
  - Pesos na **Distância Euclideana** ponderada determina a importância de uma variável.
  - **Pesos** próximos a **zero** significa que o atributo não é usado.
- **Redes Bayesianas**:
  - Estatísticas demonstram que **algumas variáveis** têm **pouco** ou **nenhum** efeito no modelo.

- Por que a gente precisa de seleção de atributos?

Motivação



Motivação ...



Motivação ...

- Seleção de variáveis (**feature selection**) tem recebido atenção especial em aplicações que usam **conjuntos de dados** com muitos atributos.
- **Exemplos**:
  - Processamento de texto.
  - Recuperação de informação em banco de imagens.
  - Bioinformática.
  - Química combinatorial.
  - etc.

Objetivos

- Os **alvos principais** do processo de **seleção de variáveis** são:
  - Melhorar a **performance** dos algoritmos de aprendizado de máquina.
  - Simplificar os **modelos de predição** e reduzir o **custo computacional** para "rodar" esses modelos.
  - Fornecer um **melhor entendimento** sobre os resultados encontrados, uma vez que existe um estudo prévio sobre o **relacionamento** entre os **atributos**.

Objetivos ...

- **IDEIA GERAL**: Processo que escolhe um **subconjunto ótimo de atributos** de acordo com uma função objetivo.
- **Objetivos**:
  - **Reduzir** dimensionalidade e **remover** ruído.
  - **Melhorar a performance** da mineração de dados:
    - Aumenta a velocidade do aprendizado.
    - Melhora a acurácia de modelos preditivos.
    - Facilita a compreensão dos resultados minerados.

Objetivos ...

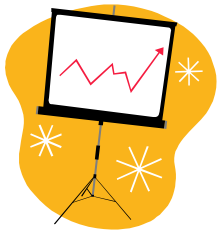
- Obter uma **representação reduzida do *dataset***, em termos de atributos, mas que produza os mesmos (**ou quase os mesmos**) resultados analíticos.
- Eliminar **atributos redundantes**:
  - **Variáveis altamente correlacionadas** não agregam informação para a construção de um modelo.
  - **Exemplo**: o **preço** de um produto e a **quantidade de imposto** pago por ele.
- Eliminar **atributos Irrelevantes**:
  - **Não contém informação útil** para o processo de mineração.
  - **Exemplo**: **RA** de um estudante é irrelevante para a tarefa de predição do **CR** (**coeficiente de rendimento**).

Métodos Supervisionados

Métodos supervisionados

- O foco é o **ranqueamento** de atributos.
- Diferentes conjuntos de atributos podem ser selecionados.
- Consideram os pontos **com a presença do atributo-meta**.
- Em algumas aplicações, se existem muitos atributos (**features**):
  - Selecionar os **top K** atributos (**scored features**).

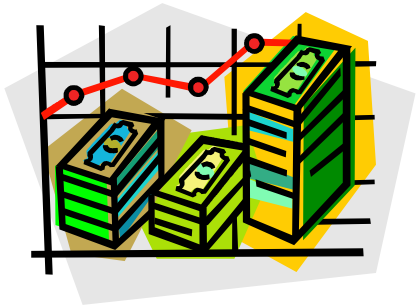
Abordagem Força Bruta



Abordagem Força-Bruta

- Tentar **todas as combinações** de atributos possíveis.
- **Ideia**: Tentar achar um subconjunto de atributos que melhor representa o conjunto original.
- Dados *N* atributos, existem 2^N subconjuntos de atributos:
  - Método **impraticável** para datasets com muitos atributos.
  - Perigo de "**overfitting**".
- **Computacionalmente proibitivo**!!

Determinação de Relevância (Embedded)



Determinação de relevância (Embedded)

- **Ideia geral**:
  - A seleção ocorre naturalmente como parte dos algoritmos de mineração.
  - Essa abordagem baseia-se no **ganho de informação**.
- **Exemplos** de algoritmos:
  - ID3;
  - C4.5 (**J48 no Weka**);
  - CART (**Classification And Regression Trees**).

Ganho de Informação

- Ranqueia os atributos através do **ganho de informação**.
  - Ganho de Informação → redução da entropia.
- O valor da **entropia** corresponde à **impureza do atributo**, a falta de homogeneidade.
- O **ganho de informação** corresponde à **variação da impureza**.
- Os **atributos** com o maior **ganho de informação** são retidos, pois ajudam a discriminar o **atributo-meta**.
- Estes atributos **minimizam** a **informação necessária** para classificar as instâncias com classes desconhecidas e refletem a **menor aleatoriedade** ou **impureza**.

Ganho de Informação ...

- Assim, a **informação esperada** (ou **entropia**) para classificar uma instância in *D* é dada por:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- onde ***pi*** é a probabilidade de que uma instância arbitrária em *D* pertença a classe C_i e seja estimada por $|C_i,D| / |D|$.
- A função ***log*** é usada na **base 2** porque a informação é codificada em bits.
- **Atributos** com **menores entropia** terão **maior ganho de informação** → devem ser selecionados.
- Essa abordagem pode se tornar **impraticável** quando o número de atributos é muito grande.

Ganho de Informação ...

- A medida **ganho de informação** tem um vício natural (**bias**) — ela favorece atributos com muito valores.
- **Por exemplo**, um atributo (**feature**) tendo diferentes valores em diferentes amostras gera uma medida (**ganho de informação**) pobre (**viciada**).
- **Solução**: Usar a taxa de ganho de informação (***information gain ratio***).
- A medida (**taxa de ganho de informação**) tenta corrigir o "**vício**" dos atributos que contêm muitos valores através da incorporação de quantidade de informação segmentada (**amount of *split information***).

Taxa de Ganho de Informação

- A **quantidade de informação segmentada** é sensível a faixa de valores de um atributo.

$$SplitInformation(f,S) = -\sum_{i=1}^{w} \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{|S|}$$

Onde:

- S_1 ... S_w são os *w* subconjuntos das amostras resultantes do particionamento de *S* pelos *w* intervalos de *f*.
- A **taxa de ganho de informação** é data por:

$$GainRation(f,S) = \frac{Gain(f,S)}{SplitInformation(f,S)}$$

Wrappers

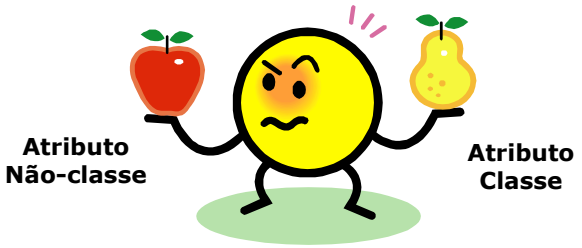
Wrappers

- **Ideia geral**:
  - Avalia conjuntos de atributos usando um **algoritmo de aprendizado de máquina**.
  - O algoritmo funciona como uma **caixa preta** para encontrar os melhores subconjuntos de atributos.
  - O propósito é encontrar o conjunto de atributos que melhor se adequa ao **algoritmo de aprendizado**.
- Essa abordagem é **totalmente dependente** do algoritmo de aprendizado.

Wrappers ...

- Os melhores atributos para o algoritmo **kNN** e **redes neurais** pode não ser os melhores para **árvores de decisão**.
- **Forward stepwise selection**:
  - Começa com um conjunto vazio ***A***. Os melhores atributos são determinados e adicionados ao conjunto ***A***.
- **Backwards elimination**:
  - Começa com um conjunto de todos os atributos. Os piores atributos são determinados e removidos do conjunto inicial.
- **Bi-directional stepwise selection & elimination**:
  - Combina as duas abordagens acima.

Qui-quadrado (χ^2)



Qui-quadrado (χ^2)

- Esse método avalia os atributos individualmente usando a medida χ^2 com relação ao **atributo-meta**.
- Quanto maior o valor de χ^2, mais provável é a **correlação das variáveis (atributo e classe)**.
- χ^2 (**teste do qui-quadrado**)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

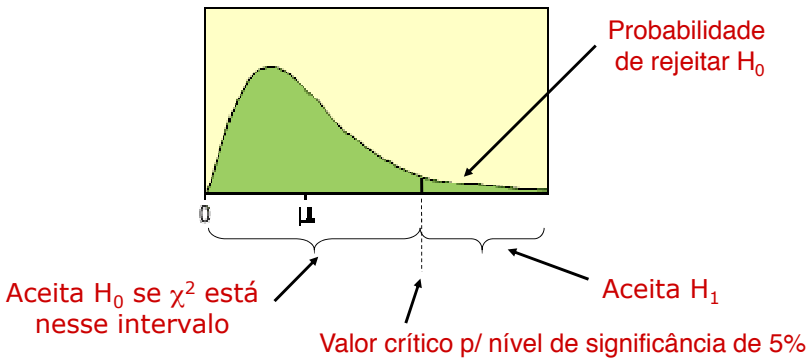
- As **frequências observadas** são obtidas diretamente dos dados das amostras, enquanto que as **frequências esperadas** são calculadas a partir destas.

Qui-quadrado (χ^2) ...

- O **analista de dados** estará sempre trabalhando com duas hipóteses:
  - H_0: não há associação entre os atributos (**independência**)
  - H_1: há associação entre os atributos.
- A hipótese H_0 é rejeitada para valores elevados de χ^2.
- O cálculo dos **graus de liberdade** de χ^2 é dado por:
  gl = (número de linhas – 1) × (número de colunas – 1)
- A perda de um grau de liberdade, isto é, o uso de *num_linhas* – **1**, deve-se ao fato de empregarmos na fórmula a média amostral em vez da média populacional.

Qui-quadrado (χ^2) ...

A forma da função de densidade de χ^2



Rejeitamos a **hipótese nula** se χ^2 for maior que o **valor crítico** fornecido pela tabela. Para 1 grau de liberdade, o valor crítico é 3,841.

Exemplo do cálculo de χ^2

	Joga xadrez	Não joga xadrez	Soma (linhas)
Gosta de ficção científica	250(90)	200(360)	450
Não gosta de ficção científica	50(210)	1000(840)	1050
Soma (colunas)	300	1200	1500

300*450 / 1500 = 90 , etc.

- **Os números entre parênteses são os valores esperados**, calculados com base na distribuição dos dados das duas categorias.
- O resultado mostra que gostar_ficção_científica e jogar_xadrez são correlacionadas nesse grupo:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Neste caso, a hipótese nula é rejeitada, pois 507.93 > 3.841. Então, existe associação entre as variáveis estudadas.

Seleção baseada em Correlação (CFS)

Seleção baseada em Correlação

- A maioria dos métodos de seleção de atributos anteriores avaliam os atributos em termos de **relevância individual** considerando as amostras em diferentes classes.
- É possível ranquear **subconjuntos de atributos**?
- ***Correlation-based feature selection*** (**CFS**) é um método em que um conjunto de atributos é considerado bom se:
  - Contém atributos altamente correlacionados com o **atributo-meta**;
  - Contém atributos não correlacionados entre si.
- O coração do método **CFS** é uma heurística de avaliação de subconjuntos que considera:
  - Não somente a utilidade de atributos individuais, mas também o nível de correlação entre eles.

Método CFS

- **CFS** primeiro calcula uma matriz de correlação de **atributo-classe** e **atributo-atributo**.
- Um peso (**score**) de um conjunto de atributos é associado usando a seguinte fórmula:

$$Mérito(S) = \frac{k \times \overline{r_{ac}}}{\sqrt{k + k(k-1)\,\overline{r_{aa}}}}$$

Onde:

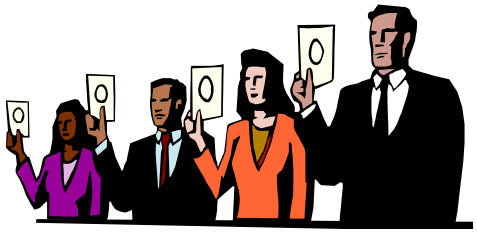
- *Mérito(S)* é o mérito de um **subconjunto de atributos** *S* contendo *k* atributos;
- $\overline{r_{ac}}$ é a média da correlação entre **atributo-classe**;
- $\overline{r_{aa}}$ é a média da correlação entre **atributo-atributo**.

Método CFS ...

$$Mérito(S) = \frac{k \times \overline{r_{ac}}}{\sqrt{k + k(k-1)\,\overline{r_{aa}}}}$$

- O numerador pode ser visto como um indicador do **poder preditivo** do conjunto de atributos.
- O denominador indica o "**grau de redundância**" que existe entre os atributos.
- **CFS** começa com o conjunto vazio de atributos e usa a heurística **best-first-search** com um critério de parada de 5 consecutivos subconjuntos que não melhoram o mérito.
- O subconjunto com o **maior mérito** encontrado pela heurística será selecionado.

Benchmark de Métodos para Seleção de Atributos



Experimentos

- **Metodologia**:
  - Avaliar a melhor abordagem de **seleção de atributos** para cada um dos **métodos de classificação** apresentados.
  - Comparar as abordagens de seleção de atributos entre si e com o conjunto original de atributos (**sem seleção**).
- **Conjuntos de Dados**:

Dataset	# Instâncias	# Atributos	# Classes
Soybean	683	36	19
Hortalicas	2000	21	3

Algoritmos

Método	Algoritmo
Árvore de Decisão	C4.5
Classificador Bayesiano	Naïve Bayes
Rede Neural	Multilayer Perceptron
Support Vector Machine	SMO

- **Software**:
  - **Weka, versão 3.6.5**.
  - **http://www.cs.waikato.ac.nz/ml/weka/**

Resultados – Dataset Soybean

Algoritmo	Sem Seleção de atributos	χ^2	InfoGain	CFS	Wrapper
C4.5	91.50	90.48	90.77	90.19	92.97
Naïve Bayes	92.97	92.82	92.97	92.24	93.11
Multilayer Perceptron	93.41	93.11	92.97	93.85	92.24
SMO	93.85	94.28	94.43	93.85	93.85

- χ^2 : atributos removidos: 5, 6, 7, 10.
- InfoGain: atributos removidos: 5, 9, 10, 25.
- CFS: atributos removidos: 2, 6, 14, 16, 20, 21, 25, 27, 29, 31, 32, 33, 34.
- Wrapper (C4.5): atributos removidos: 2, 6, 7, 8, 9, 10, 12, 16, 28, 32.

Resultados – Dataset Hortaliças

Algoritmo	Sem Seleção de atributos	χ^2	InfoGain	CFS	Wrapper
C4.5	90.75	89.91	90.75	94.11	94.11
Naïve Bayes	72.54	77.03	75.35	60.06	73.10
Multilayer Perceptron	82.35	91.59	90.75	66.10	92.43
SMO	82.07	80.95	80.95	61.06	80.39

- χ^2 : atributos removidos: 3, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17.
- InfoGain: atributos removidos: 3, 6, 9, 10, 14, 15, 17.
- CFS: atributos selecionados: 5, 11, 18, 19.
- Wrapper (C4.5): atributos selecionados: 3, 4, 5, 7, 18, 19.

Ranqueamento de Variáveis

Ranqueamento de Variáveis

- **Cuidado!!**
- Redução de ruído e melhor precisão de uma classificação podem ser obtidos **adicionando-se variáveis** que presumidamente são redundantes.
- **Variáveis perfeitamente correlacionadas** são redundantes no sentido de que não agregam nenhuma informação se forem adicionadas ao modelo.
- **Duas variáveis** que **não** têm grande significado **isoladas** podem ser úteis quando estão **juntas** em um modelo.

Seleção de Atributos: Aspectos Relevantes

- A **Seleção de atributos** (**SA**) quase sempre melhora a precisão de modelos em problemas reais.
- **Aspectos relevantes sobre SA**:
  - Simplifica modelos;
  - Torna os modelos mais inteligíveis;
  - Ajuda a explicar melhor um problema real;
  - Evita o problema: "**Princípio de Economia Científica**"

Princípio de Economia Científica:

"Quanto menos se sabe a respeito de um fenômeno, maior o número de variáveis exigidas para explicá-lo"

Seleção de Atributos: Limitações

- Considerando um *dataset* com muitos atributos, a seleção de atributos pode causar ***overfit***.
- ***Wrappers*** requerem que os algoritmos de aprendizado rodem muitas vezes, o que é **muito caro**!
- Quando um **atributo não é selecionado**, não significa que esse atributo não é importante.
- Alguns **atributos descartados** podem ser muito **importantes para especialistas** do domínio.
- Muitos dos métodos são **gulosos** e **não** trabalham com otimização do conjunto de atributos selecionados.

Seleção de Atributos: Desafios

- Heurísticas para acelerar o processo de seleção de atributos (**Exemplo: 1000 atributos**).
- Métodos para prevenir ***overfitting***.
- Métodos para **selecionar atributos relevantes** sem depender dos algoritmos de aprendizado de máquina.
- **Detecção de Irrelevância**:
  - Atributos realmente irrelevantes podem ser ignorados;
  - Melhores algoritmos;
  - Melhores definições para formulação de heurísticas.

Referências para consulta

- **JMLR Special Issue on Variable and Feature Selection**. Disponível em **http://jmlr.csail.mit.edu/papers/special/feature03.html**
- J. T. Tou; R. C. Gonzalez. **Pattern Recognition Principles**. Addison-Wesley, 1974.
- Lui,H and Setiono,R. (1996). **Feature selection and classification - a probabilistic wrapper approach**. In *Proceedings of the 9th Intl. Conf. on Industrial and Engineering Applications of AI and ES*.
- Kohavi, R., and Sommerfield, D. (1995). **Feature subset selection using the wrapper model**: Overfitting and dynamic search space topology. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95).*
- M.A. Hall and G. Holmes. **Benchmarking attribute selection techniques for discrete class data mining**. *IEEE Transaction on Knowledge and Data Engineering*, 15(3):in press,May/June 2003.

Referências para consulta ...

- M.A. Hall. ***Correlation-based feature selection for machine learning***. PhD thesis, Department of Computer Science, University of Waikato, Hamilto, New Zealand, 1998.
- U. Fayyad and K. Irani. **Multi-interval discretization of continuous-valued attributes for classification learning**. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029, 1993.
- H. Liu and R. Setiono. **Chi2**: **Feature selection and discretization of numeric attributes**. *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, pages 388–391, November 1995.
- T.M. Mitchell. ***Machine Learning***. McGrawHill, USA, 1997.
- P.J. Park, M. Pagano, and M. Bonetti. **A non-parametric scoring algorithm for identifying informative genes from microarray data**. *Pacific Symposium on Biocomputing*, pages 52–63, 2001.