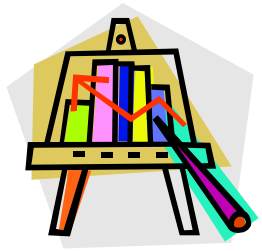
Métodos de Seleção de Atributos para Mineração de Dados

Stanley Robson de M. Oliveira

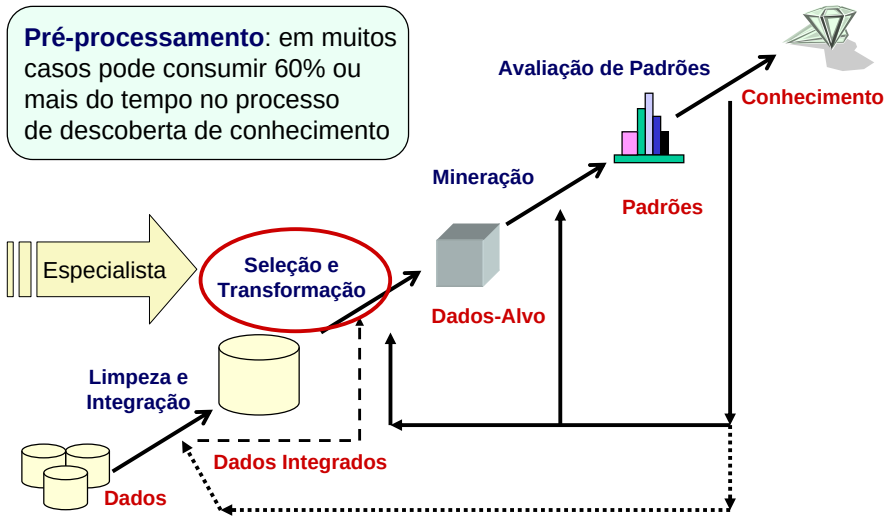


Agenda

- **Seleção de atributos**:
  - Motivação e objetivos.
- **Abordagens para seleção de atributos**:
  - Métodos Supervisionados.
- **Estudo de caso**
  - Comparação de métodos supervisionados.
- **Conclusão**:
  - Aspectos relevantes;
  - Limitações;
  - Desafios de pesquisa.

O processo de descoberta do conhecimento

Pré-processamento: em muitos casos pode consumir 60% ou mais do tempo no processo de descoberta de conhecimento



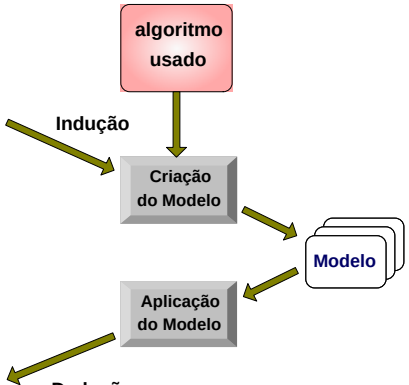
Geração de um Modelo de Classificação

Tid	Atrib1	Atrib2	Atrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de treinamento

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

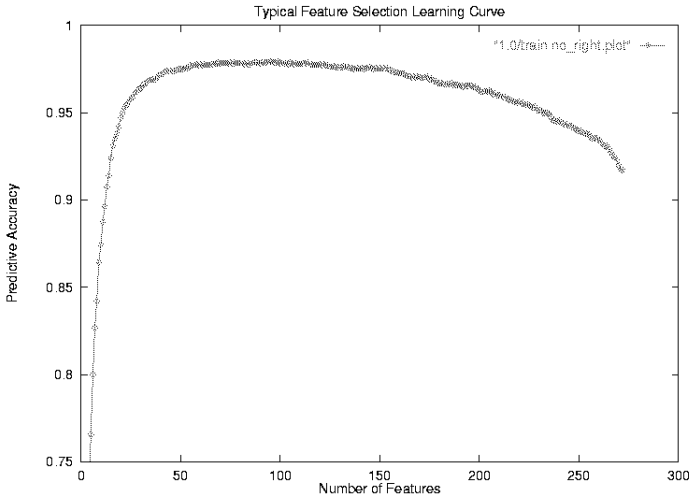
Conjunto de Teste



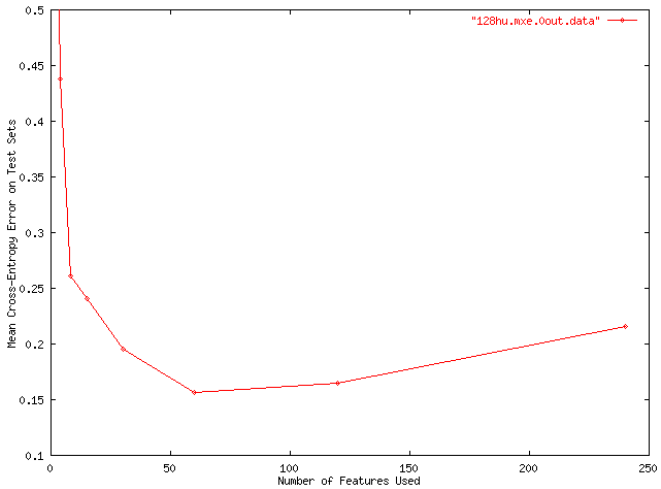
Classificação de Dados

- **Classificação**: uma tarefa **preditiva**: define o valor de uma variável desconhecida (**atributo classe**) a partir de variáveis conhecidas.
- **Aplicações**:
  - Classificar tumores como benigno ou maligno.
  - Detecção de fraudes;
  - Diagnósticos médicos;
  - Avaliação de riscos de empréstimos.
  - Etc.

Motivação



Motivação ...



Motivação ...

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

Objetivos

- Os **alvos principais** do proceso 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 ...

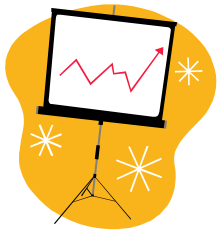
- Obter uma **representação reduzida do *dataset***, em termos de atributos, mas que produza os mesmos (**ou quase os mesmos**) resultados analíticos.
- Eliminar **atributos redundantes**:
  - **Variáveis altamente correlacionadas** não agregam informação para a construção de um modelo.
  - **Exemplo**: 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**: ID de um estudante é irrelevante para a tarefa de predição do GPA (**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 influência do atributo-classe**.
- Em algumas aplicações, se existem muitos atributos (**features**):
  - Selecionar os **top K** atributos (**scored features**).

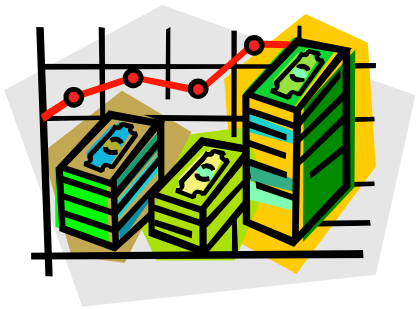
Abordagem Força Bruta



Abordagem Força-Bruta

- Tentar **todas as combinações** de atributos possíveis.
- **Idéia**: 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)

- **Idéia 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.

Ganho de Informação

- Ranqueia os atributos através do **ganho de informação**.
  - Ganho de Informação → redução da entropia.

$$Ent(S) = \sum_{i=1}^{k} -p_i \times \log_2(p_i) \qquad Gain(f,S) = Ent(S) - Ent(f,T_f,S)$$

$$Gain(f,S) = Ent(S) - \sum_{v \in Values(f)} \frac{|S_v|}{|S|} \times Ent(S_v)$$

- Tenta os 10, 20, 30, …, N atributos no aprendizado.
- Avalia por meio do conjunto de testes (ou **validação cruzada**).
- Essa abordagem se torna **impraticável** quando o número de atributos é muito grande.

Wrappers

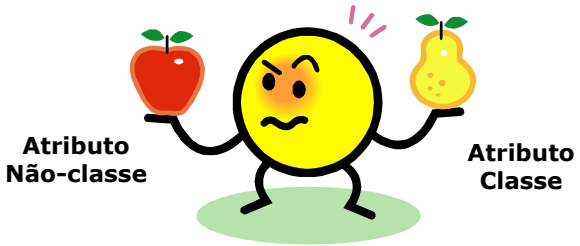
Wrappers

- **Idéia 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 à classe**.
- 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 **freqüências observadas** são obtidas diretamente dos dados das amostras, enquanto que as **freqüências esperadas** são calculadas a partir destas.

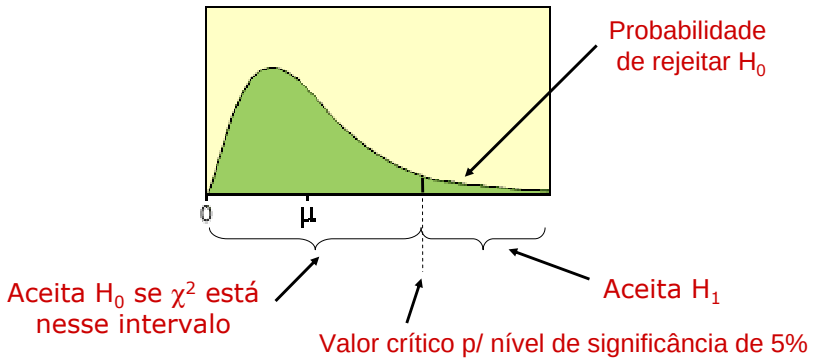
Qui-quadrado (χ^2) ...

- 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.
- Um atributo f_a se torna **mais relevante** do que um atributo f_b , com ($a \neq b$) se $\chi^2(f_a) > \chi^2(f_b)$.
- O cálculo dos **graus de liberdade** de χ^2 é dado por:

  gl = (número de linhas – 1) x (número de colunas – 1).

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

- **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, as variáveis estudadas são correlacionadas.

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 a classe;
  - 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:

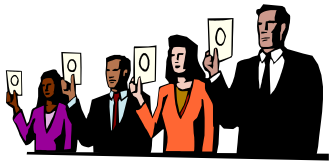
- *Mérito(S)* é o mérito de um **subconjunto de atributos** *S* contendo *k* atributos;
- $\overline{r_{ac}}$ é a média da correlação entre **atributo-classe**;
- $\overline{r_{aa}}$ é a média da correlação entre **atributo-atributo**.

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



Experimentos

- **Metodologia**:
  - Avaliar a melhor **abordagem de seleção de atributos** para alguns **métodos de classificação** existentes.
  - 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 3.4.8**.
  - **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 Hortalicas

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.

Conclusões

Ranqueamento de Variáveis

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 do "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.

Obrigado pela atenção !!

Stanley Robson de Medeiros Oliveira
`stanley@cnptia.embrapa.br`

Embrapa Informática Agropecuária
Av. André Tosello, 209
Caixa Postal 5041, Campinas, SP

Referências para consulta

- Wall, Michael E., Andreas Rechtsteiner, Luis M. Rocha. **Singular value decomposition and principal component analysis**. In *A Practical Approach to Microarray Data Analysis*. D.P. Berrar, W. Dubitzky, M. Granzow, eds. pp. 91-109, Kluwer: Norwell, MA, 2003.
- Papadimitriou CH, Tamaki H, Raghavan P, Vempala S. **Latent semantic indexing: a probabilistic analysis**. In: Proceedings of the 17th ACM symposium on principles of database systems. Seattle, WA, USA; June 1998. p. 159–68.
- Jolliffe, I. T. **Discarding Variables in a Principal Component Analysis.** In *Applied Statistics*, Vol. 21, No. 2 (1972), pp. 160-173.
- Jolliffe, I. T. **Principal Component Analysis**: Springer-Verlag, New York, 1986.

Referências para consulta ...

- Kaski S. **Dimensionality reduction by random mapping**. In: Proceedings of the international joint conference on neural networks. Anchorage, Alaska; May 1999. p. 413–18.
- Kruskal JB, Wish M. **Multidimensional scaling**. Beverly Hills, CA, USA: Sage Publications; 1978.
- Larsen B, Aone C. **Fast and effective text mining using linear-time document clustering**. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. San Diego, CA, USA; August 1999. p. 16–22.
- Faloutsos C, Lin K-I. **FastMap: a fast algorithm for indexing, datamining and visualization of traditional and multimedia datasets**. In: Proceedings of the 1995 ACM SIGMOD international conference on management of data. San Jose, CA, USA; June 1995. p. 163–74.

Referências para consulta ...

- Bingham E, Mannila H. **Random projection in dimensionality reduction: applications to image and text data**. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA, USA; 2001. p. 245–50.
- Johnson WB, Lindenstrauss J. **Extensions of Lipshitz mapping into Hilbert space**. In: Proceedings of the conference in modern analysis and probability. Contemporary mathematics, vol. 26; 1984. p. 189–206.
- Achlioptas D. **Database-friendly random projections**. In: Proceedings of the 20th ACM symposium on principles of database systems. Santa Barbara, CA, USA; May 2001. p. 274–81.
- Fern XZ, Brodley CE. **Random projection for high dimensional data clustering: a cluster ensemble approach**. In: Proceedings of the 20th international conference on machine learning (ICML 2003). Washington DC, USA; August 2003

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.

Referências para consulta ...

- R. Sandy. ***Statistics for Business and Economics***. McGrawHill, USA, 1989.
- F. Wilcoxon. **Individual comparisons by ranking methods**. *Biometrics*, 1:80–83, 1945.
- E.P. Xing and R.M. Karp. **Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts**. *Proceedings of The Ninth International Conference on Intelligence Systems for Molecular Biology, published on Bioinformatics*, 17(suppl):S306–S315, 2001.
- Kenney, J. F. and Keeping, E. S. ***Mathematics of Statistics****, Pt. 2, 2nd ed.* Princeton, NJ: Van Nostrand, 1951.
- Weisstein, Eric W. **Chi-Squared Test**. From *MathWorld* – A Wolfram Web Resource. http://mathworld.wolfram.com/Chi-SquaredTest.html