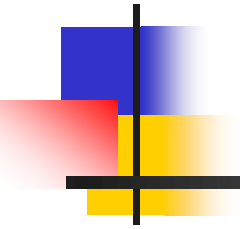
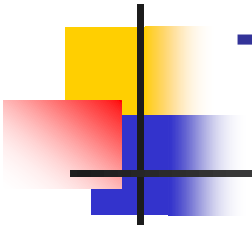


SCC5895 – Análise de Agrupamento de Dados

Apresentação do Curso

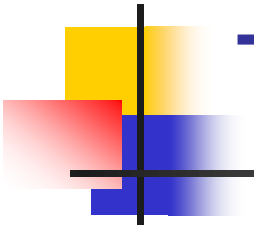
Prof. Eduardo Raul Hruschka

PPG-CCMC / ICMC / USP



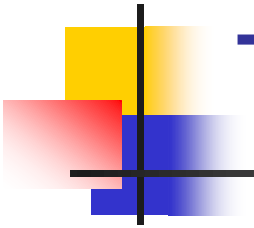
Tópicos do Curso

- **Conceituação de Análise de Agrupamento**
  - Conceitos e definições básicas, motivação, aplicações, ...
- **Representação de Dados**
  - Tipos de dados, normalizações, medidas de similaridade e dissimilaridade, ...
- **Métodos Hierárquicos**
  - Métodos aglomerativos, métodos divisivos, formulações parametrizadas, relação com teoria dos grafos, ...



Tópicos do Curso

- **Métodos Particionais**
  - Métodos de partições rígidas (baseados em protótipos e densidade), métodos de partições com sobreposição (*fuzzy* e probabilística), ...
- **Avaliação de Agrupamentos**
  - Índices de validação externos, internos e relativos (para partições rígidas e com sobreposição), técnicas de validação de estruturas hierárquicas e particionais, qualidade de grupos individuais, tendência de agrupamento, estimação do número de grupos, ...



Tópicos do Curso

- **Eventuais Tópicos Avançados**
  - Técnicas de agrupamento paralelo e distribuído de dados
  - Métodos estatísticos para comparação de algoritmos
  - Estruturas de dados para escalabilidade computacional de algoritmos
  - Análise de estabilidade de agrupamento via re-amostragem (resampling)
  - Agrupamento de fluxo de dados (data streams)
  - Agrupamento de dados em sub-espaços (biclustering, redução de atributos, ...)
  - Combinações de agrupamentos (clustering ensembles)
  - Agrupamento semi-supervisionado
  - Agrupamento de sequências
  - ...

Calendário

Aula	Data	Assunto / Atividade
1	09/08	Apresentação da Disciplina / Introdução
2	16/08	Representação de Dados. Medidas de Proximidade.
3	23/08	Medidas de Proximidade. Conversão de Escalas.
4	30/08	Métodos Hierárquicos.
	06/09	Recesso.
5	13/09	Métodos Hierárquicos.
6	20/09	Prova 1.
7	27/09	Métodos Particionais.
8	04/10	Métodos Particionais.
9	11/10	Validação.
10	18/10	Validação.
11	25/10	Validação.
12	01/11	Prova 2.
	08/11	Palestra – Bragfost (não há aula).
	15/11	Feriado.
13	22/11	Tópicos Avançados - Apresentação de Seminários.
14	29/11	Apresentação de Projetos.
15	06/12	Apresentação de Seminários/Projetos e Discussões.

Avaliação

- A avaliação será composta por:

  - Duas provas ($P_1 \in [0,10]$, $P_2 \in [0,10]$);

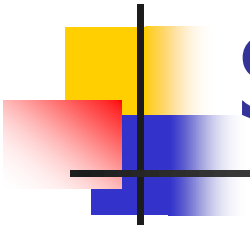
  - Um projeto de curso $P_C \in [0,10]$ (implementação/experimentação);

  - Uma apresentação de seminário $S \in [0,10]$ (tópico avançado/artigo recente);

- As notas obtidas serão consolidadas em uma única nota final:

  $NF = (P_1+P_2+P_C+S)/4$

- Aprovação: C[5,7] B(7,8.5] A(8.5,10] (Frequência ≥ 75%).



Seminários

- Deverão abordar tópicos / algoritmos mais avançados não vistos em aula ou que estendam o conteúdo visto em aula. Em particular:

1. Latent Dirichlet Allocation (LDA) e extensões
2. K-means eficiente (árvores kd e desigualdade triangular)
3. DBSCAN eficiente baseado em árvores R* ou kd
4. CHAMELEON (Grafo / Hierárquico)
5. CLIQUE (Subespaço / Grid / Densidade)
6. BIRCH (CF-trees / Acesso Sequencial / Grandes BDs)
7. OPTICS (Densidade / Visualização)
8. CURE (Amostragem / Múltiplos Representativos)
9. COP-kmeans (Semi-Supervisionado)
10. CLICK (Grafo / Probabilístico)

11. DENCLUE (Densidade)
12. Agrupamento HMM (Seqüências / Dados Temporais)
13. MAFIA (Subespaço / Grid / Densidade)
14. ROCK (Grafo / Atributos Categóricos)
15. DBSCAN Incremental (Acesso Sequencial / Grandes BDs)
16. Agrupamento via Acúmulo de Evidência (Ensemble)
17. PROCLUS ou P3C (Projeção / Subespaço)
18. SUBCLU ou STATPC (Subespaço / Dens. / Estatística)
19. CLARANS (Busca / Medóides)
20. X-Means

Bibliografia Principal

- Jain, A. K., Dubes, R. C., **Algorithms for Clustering Data**, Prentice Hall, 1988.
- Xu, R., Wunsch, D., Clustering, IEEE Press, 2009.
- Gan, G., Ma, C., Wu, J., Data Clustering: Theory, Algorithms, and Applications, SIAM Series on Statistics and Applied Probability, 2007.
- Kogan, J., Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, 2006.
- Everitt, B. S., Landau, S., Leese, M., Cluster Analysis, Hodder Arnold Publication, 2001.
- Arabie, P., Hubert, L. J., Soete, G., Clustering and Classification, World Scientific Publ., 1996.
- Höppner, F., Klawonn, F., Kruse, R., Runkler, T., Fuzzy Cluster Analysis, 1999.
- Kaufman, L., Rousseeuw, P. J., Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, 1990.
- Hartigan, J. A., Clustering Algorithms, John Wiley & Sons, 1975.
- Anderberg, M. R., Cluster Analysis for Applications, Academic Press, 1973.
- …
- **Artigos de periódicos especializados**…