



SCC5895 – Análise de Agrupamento de Dados

Apresentação do Curso

Prof. Ricardo J. G. B. Campello

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, ...

3



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-espacos (biclustering, redução de atributos, ...)
 - Combinações de agrupamentos (clustering ensembles)
 - Agrupamento semi-supervisionado
 - ...

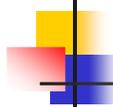


Calendário (Sujeito a Alterações)

- | | |
|--|---|
| <ul style="list-style-type: none"> ▪ 12/08 – Aula ▪ 19/08 – Aula ▪ 26/08 – Aula ▪ 02/09 – Aula ▪ 16/09 – Aula ▪ 23/09 – PROVA I ▪ 30/09 – Aula ▪ 07/10 – Aula | <ul style="list-style-type: none"> ▪ 14/10 – Aula ▪ 21/10 – Aula ▪ 11/11 – Aula ▪ 18/11 – PROVA II ▪ 25/11 – Seminários ▪ 02/12 – Seminários (*) ▪ 09/12 – Seminários |
|--|---|

(*) Data limite para entrega do projeto

➤ Eventuais alterações serão comunicadas na webpage do curso



Avaliação

- **Nota Final:** $NF = (w_1 * N_{pv} + w_2 * N_{pj} + w_3 * N_s) / (w_1 + w_2 + w_3)$
 - onde N_{pv} , N_{pj} e N_s são as notas de prova, projeto e seminário
 - a nota de prova (N_{pv}) é uma média aritmética das notas da PI e PII
 - os pesos de cada uma das três notas que compõem a média ponderada são dados da seguinte maneira:
 - $w_i = 10 - \min(5, \text{Nota})$
 - logo, NF se torna uma média aritmética caso nenhuma nota seja < 5



Seminários

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

- | | |
|---|--|
| 1. K-means eficiente baseado em árvores kd | 11. Agrupamento HMM (Dados Sequenciais / Temporais) |
| 2. K-means eficiente baseado em des. triangular | 12. MAFIA (Subespaço / Grid / Densidade) |
| 3. DBSCAN eficiente baseado em árvores kd | 13. CLARANS (Busca / Medóides) |
| 4. CHAMELON (Grafo / Hierárquico) | 14. DENCLUE (Densidade) |
| 5. CLIQUE (Subespaço / Grid / Densidade) | 15. PROCLUS ou P3C (Projeção / Subespaço) |
| 6. BIRCH (CF-trees) | 16. ROCK (Grafo / Atributos Categóricos) |
| 7. CURE (Amostragem) | 17. Agrupamento via Acúmulo de Evidência (Ensemble) |
| 8. COP-kmeans (Semi-Supervisionado) | 18. X-Means |
| 9. CLICK (Grafo) | 19. SUBCLU ou STATPC (Subespaço / Dens. / Estatística) |
| 10. OPTICS (Densidade) | |



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, 1999.
- 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...**