# Universidade de Pernambuco
# Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

## Proposta de Dissertação de Mestrado

**Área: Inteligência Computacional**

**Título: Desenvolvimento de Técnica de Clusterização utilizando Teoria de Informação e Inteligência Computacional**

**Orientador – Carmelo José Albanez Bastos Filho (carmelofilho@upe.br)**

**Co-orientador – Diego Pinheiro (pinsilva@ucdavis.edu)** - University of California, Davis

**Descrição**

Algoritmos de clusterização encontram agrupamentos em diversos dados clínicos e moleculares, podendo levar a criação de terapias mais personalizadas [1], como socioeconômicos, podendo levar a à criação de políticas de saúde pública [2]. Questões metodológicas, no entanto, ainda carecem de maior rigor científico como, por exemplo, a simples determinação do número de agrupamentos existentes em um determinado banco de dados [1-5]. Desde a criação do primeiro algoritmo de clusterização [3], novas metodologias têm sido propostas, envolvendo tanto a criação de novos algoritmos para encontrar agrupamentos, quanto a proposição de novas métricas para validar a qualidade dos agrupamentos encontrados de maneira não supervisionada [4-5]. Como os algoritmos de clusterização tendem a buscar por agrupamentos que otimizem determinadas métricas de validação interna, as diferentes métricas de validação interna frequentemente discordam entre si quanto à qualidade dos agrupamentos encontrados até por um mesmo algoritmo num mesmo banco de dados [4]. Essas métricas são geralmente baseadas em premissas irrealistas pois assumem que os dados apresentam propriedades estatísticas consideravelmente mais simples que as comumente exibidas em dados reais. Independentemente, tais métricas de validação interna, em conjunto com variados algoritmos de clusterização, têm sido aplicadas em dados reais para validar agrupamentos que potencialmente impactarão diretamente a população [1-2]. Para superar os desafios metodológicos em clusterização, novas abordagens baseadas em teoria de informação [5] e inteligência computacional [6] redefiniram o modelo de agrupamentos encontrados e do processo para encontrar estes agrupamentos, respectivamente. Atualmente, o grande desafio em clusterização é a ausência de uma caracterização do aprendizado não supervisionado durante o processo de clusterização por meio de métricas de validação interna utilizando premissas mais realistas sobre os dados. Recentemente, um esforço envolvendo a UC Davis (EUA) e UPE (Brasil) propôs uma abordagem baseada em teoria de informação que rastreia o ganho de informação durante o processo de clusterização para caracterizar o processo que levou ao melhor aprendizado não supervisionado. Foi mostrado que tal abordagem pode, por exemplo, levar sistematicamente à escolha do número de agrupamentos mais adequado principalmente em dados reais. O objetivo deste trabalho de dissertação de mestrado é mostrar que a utilização de *teoria de informação e inteligência computacional* na caracterização dos agrupamentos encontrados e no processo de busca por tais agrupamentos, respectivamente, pode levar a um melhor entendimento do *aprendizado não supervisionado durante o processo de clusterização* e, por fim, ao desenvolvimento de uma nova técnica de clusterização mais adequada para dados reais com propriedades estatísticas mais complexas. A validação da proposta será realizada em bases de dados de saúde.

**Referências Bibliográficas**

1. Seymour, C. W., Seymour, C. W., Kennedy, J. N., Kennedy, J. N., Wang, S., Wang, S., et al. (2019). Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. Jama, 1–15. http://doi.org/10.1001/jama.2019.5791
2. Wallace, M., Sharfstein, J. M., Kaminsky, J., & Lessler, J. (2019). Comparison of US County-Level Public Health Performance Rankings With County Cluster and National Rankings. JAMA Network Open, 2(1), e186816–11. http://doi.org/10.1001/jamanetworkopen.2018.6816
3. Sorensen, T. J. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Bilogiske Skrifter, 5, 1–35.
4. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411–423. http://doi.org/10.1111/1467-9868.00293
5. Agnelli, J. P., Cadeiras, M., Tabak, E. G., Turner, C. V., & Vanden-Eijnden, E. (2010). Clustering and Classification through Normalizing Flows in Feature Space. *Multiscale Modeling & Simulation*, *8*(5), 1784–1802. http://doi.org/10.1137/100783522
6. Figueiredo, E., Macedo, M., Siqueira, H. V., Santana, C. J., Jr, Gokhale, A., & Bastos-Filho, C. J. A. (2019). Swarm intelligence for clustering --- A systematic review with new perspectives on data mining. Eng. Applications of Artificial

Intelligence, 82, 313–329. http://doi.org/10.1016/j.engappai.2019.04.007