Fichamento a respeito de algoritmos de clusterização para aprendizado de maquina

Antônio Cícero Amorim de Azevedo - 811455 - Graduação
Docente Responsável - Alexandre Levada
Ciência da Computação - Universidade Federal de São Carlos

August 5, 2024

1 Tema de Pesquisa

Este fichamento aborda o estudo de artigos científicos sobre algoritmos de aprendizado de máquina baseados em *clustering*. O objetivo é entender e resumir o que as principais revistas andam publicando a respeito de algoritmos de clusterização.

2 Artigos

O principal motor de busca foi o *Google Acadêmico* nele fiz um filtro para buscar artigos entre os anos 2019 - 2024 com as palavras chaves: *clustering, algorithms, machine learning, artificial intelligence*. Nele encontrei artigos publicadas na revista *IEEE* (2) e (3), além de um artigo publicado na *Plos one* (1)

3 Fichas catalográficas

3.1 Clustering algorithms: A comparative approach (1)

- Abstração do artigo:

> Many real-world systems can be studied in terms of pattern recognition tasks, so that proper use (and understanding) of machine learning methods in practical applications becomes essential. While many classification methods have been proposed, there is no consensus on which methods are more suitable for a given dataset. As a consequence, it is important to comprehensively compare methods in many possible scenarios. In this context, we performed a systematic comparison of 9 well-known clustering methods available in the R language assuming normally distributed data. In order to account for the many possible variations of data, we consid-

ered artificial datasets with several tunable properties (numberof classes, separation between classes, etc). In addition, we also evaluated the sensitivity ofthe clustering methods with regard to their parameters configuration. The results revealedthat, when considering the default configurations of the adopted methods, the spectralapproach tended to present particularly good performance. We also found that the defaultconfiguration of the adopted implementations was not always accurate. In these cases, a simple approach based on random selection of parameters values proved to be a good alternative to improve the performance. All in all, the reported approach provides subsidies guiding the choice of clustering algorithms.

- Link para o material completo: Clustering algorithms: A comparative approach

3.1.1 Sintese do Trabalho

- Problema tratado: Comparação sistemática de 9 métodos de *clustering* conhecidos para determinar quais métodos são mais adequados para diferentes conjuntos de dados.

- Hipótese: Ela é subentendida onde os diferentes métodos de *clustering* terão desempenhos variados dependendo dos parâmetros utilizados e das características dos conjuntos de dados.

- Métodos de pesquisa:

  - Abordagem: Quantitativa

  - Natureza: Aplicada

  - Objetivo: Comparativa

  - Procedimento: Utilização de conjunto de dados artificiais com diferentes propriedades, como número de classes e espaçamento entre classes avaliando sempre a sensibilidade dos métodos de *clustering* em relação à configuração dos seus parâmetros a partir de métricas conhecidas da clusterização e analise estatística.

- Resultados: A abordagem espectral tende a apresentar um desempenho geral bom. Também foi encontrado que a configuração padrão das implementações adotadas nem sempre foi precisa, além do mais alguns algoritmos se adaptam melhor um determinado conjunto de dados.

3.2 Survey of state-of-the-art mixed data clustering algorithms (2)

- Abstração do artigo:

> Mixed data comprises both numeric and categorical features, and mixed datasets occur frequently in many domains, such as health, finance, and marketing. Clustering is often applied to mixed datasets to find structures and to group similar objects for further analysis. However, clustering mixed data are challenging because it is difficult to directly apply mathematical operations, such as summation or averaging, to the feature values of these datasets. In this paper, we present a taxonomy for the study of mixed data clustering algorithms by identifying five major research themes. We then present the state-of-the-art review of the research works within each research theme. We analyze the strengths and weaknesses of these methods with pointers for future research directions. At last, we present an in-depth analysis of the overall challengesin this field, highlight open research questions, and discuss guidelines to make progress in the field.

- Link para o material completo: Survey of state-of-the-art mixed data clustering algorithms

3.2.1 Sintese do Trabalho

- Problema tratado: A necessidade de algoritmos de *clustering* que possam lidar eficientemente com dados mistos, ou seja, dados que contêm tanto atributos numéricos quanto categóricos.

- Hipótese: O que é subentendido é que os algoritmos de *clustering* existentes podem

ser aprimorados ou combinados para lidar melhor com dados mistos.

- Métodos de pesquisa:
  - Abordagem: Quantitativa
  - Natureza: Teórica e Aplicada
  - Objetivo: Revisão e síntese de métodos existentes
  - Procedimento:Uma revisão extensa da literatura para identificar e categorizar algoritmos de *clustering* destinados a dados mistos. Os algoritmos foram classificados de acordo com suas metodologias, como distância, modelo e densidade. Em seguida, foi realizada uma análise comparativa para discutir os pontos fortes e fracos de cada abordagem.
- Resultados: O artigo sugere que não existe um único algoritmo que seja ideal para todas as situações de *clustering* de dados mistos. Cada abordagem possui seus pontos fortes e fracos, dependendo das características dos dados e dos objetivos específicos da análise. A revisão identifica áreas onde mais pesquisas são necessárias, incluindo a melhoria da eficiência computacional, a gestão de dados de alta dimensionalidade e o desenvolvimento de métodos que possam lidar com dados ruidosos e esparsos.

3.3 Unsupervised K-means clustering algorithm (3)

- Abstração do artigo:

  > The k-means algorithm is generally the most known and used clustering method. There are various extensions of k-means to be proposed in the literature. Although it is an unsupervised learning to clustering in pattern recognition and machine learning, the k-means algorithm and its extensions are always influenced by initializations with a necessary number of clusters a priori. That is, the k-means algorithm is not exactly an unsupervised clustering method. In this paper, we construct an unsupervised learning schema for the k-means algorithm so that it is free of initializations without parameter selection and can also simultaneously find an optimal number of clusters. That is, we propose a novel unsupervised k-means (U-k-means) clustering algorithm with automatically finding an optimal number of clusters without giving any initialization and parameter selection. The computational complexity of the proposed U-k-means clustering algorithm is also analyzed. Comparisons between the proposed U-k-means and other existing methods are made. Experimental results and comparisons actually demonstrate these good aspects of the proposed U-k-means clustering algorithm.

- Link para o material completo: Unsupervised K-means clustering algorithm

3.3.1 Sintese do Trabalho

- Problema tratado: Uma limitação dos algoritmos *k-means* tradicionais, que requerem a inicialização com um número de *clusters* a priori e a influência significativa dessa inicialização no desempenho do algoritmo.

- Hipótese: É subentendido que é possível desenvolver um algoritmo de *clustering k-means* que seja completamente não supervisionado, eliminando a necessidade de inicialização e seleção de parâmetros, e que possa encontrar automaticamente o número ótimo de *clusters*. Além disso, o algoritmo *U-k-means* é capaz de ter um desempenho equivalente aos métodos existentes.

- Métodos de pesquisa:

  - Abordagem: Quantitativa

  - Natureza: Aplicada

  - Objetivo: Desenvolvimento e comparação de métodos

  - Procedimento: O desenvolvimento do algoritmo *U-k-means* foi desenvolvido para ser capaz de determinar o numero ideal de *clusters* incorporando um calculo de penalidade/entropia capaz de realizar essa ideia. Para avaliar sua eficácia, foram realizados experimentos com conjuntos de dados numéricos e reais, comparando o *U-k-means* com técnicas existentes como *k-means*. A precisão do *clustering* foi avaliada através de métricas e analises estatísticas.

- Resultados: O artigo sugere que não existe um único algoritmo que seja ideal para todas as situações de *clustering* de dados mistos. Cada abordagem possui seus pontos fortes e fracos, dependendo das características dos dados e dos objetivos específicos da análise. A revisão identifica áreas onde mais pesquisas são necessárias, incluindo a melhoria da eficiência computacional, a gestão de dados de alta dimensionalidade e o desenvolvimento de métodos que possam lidar com dados ruidosos e esparsos.

4 Pontos de Interesse

Esse tema é a linha de pesquisa na Iniciação Cientifica com o Professor Alexandre Levada, onde nos estamos trabalhando em desenvolver uma novo método de clusterização onde a forma de agrupamento se baseia nas ideias do algoritmo *kmedias* e o calculo de distancias geodésicas

References

[1] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, 14(1), e0210236.

[2] Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7, 31883–31902.

[3] Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, 8, 80716–80727.