UM MÉTODO PARA AVALIAÇÃO DA VARIABILIDADE DE CONJUNTOS DE PARTIÇÕES

Data de aceite: 01/09/2023

Nayara G. Ribeiro
Graduação em Ciência da Computação
Universidade Federal de Uberlândia – Uberlândia, MG – Brasil

Daniel D. Abdala

RESUMO: Este trabalho tem como objetivo produzir um algoritmo para geração automática de conjuntos de partições para servir como etapa de geração de dados em processos de agrupamento de dados via combinação de partições. O algoritmo visa gerar um conjunto de partições utilizando diferentes algoritmos de agrupamento de dados, utilização de parametrizações distintas no processo de agrupamento, com o intuito de gerar o conjunto de partições o mais variável possível. Resultados experimentais demostram a viabilidade da proposta.

PALAVRAS-CHAVE: Ensemble Clustering, variabilidade, agrupamento

1 | INTRODUÇÃO, MOTIVAÇÃO E FUNDAMENTAÇÃO TEÓRICA

Ensemble clustering (combinação de partições) surgiu como uma opção de agrupamento de dados. Esta técnica é uma maneira de lidar com o problema da escolha do algoritmo de agrupamento em casos em que pouco ou nada se sabe sobre o conjunto de dados [Abdala 2010]. Ele também suaviza o resultado final quando partições diferentes apresentam distribuições consideravelmente distintas.

Neste contexto, este trabalho tem como objetivo desenvolver um método para análise de variabilidade intra-partições e recomendação de partições para fins de combinação de modo que a variabilidade seja maximizada ou minimizada.

Para realizar o agrupamento de dados, utilizamos algoritmos aproximados. Atualmente, há uma infinidade de algoritmos aproximados, que utilizam diferentes heurísticas para encontrar resultados sub ótimos. Estes algoritmos são desenvolvidos especialmente para conjuntos de dados específicos. Dado um novo conjunto a ser agrupado, não se sabe, a priori, qual o melhor algoritmo a ser aplicado.

A utilização da técnica de clustering determina o agrupamento intrínseco em conjuntos de dados. O processo de organização dos dados surge da semelhança entre os dados de alguma forma. Um cluster é, portanto, coleções de dados que são semelhantes entre eles, e que são diferentes dos dados pertencentes a outros clusters.

As etapas que serão seguidas para o desenvolvimento de um trabalho de seleção de partições que priorize aquelas que apresentam maior variabilidade entre si, podem ser observadas na figura 1.

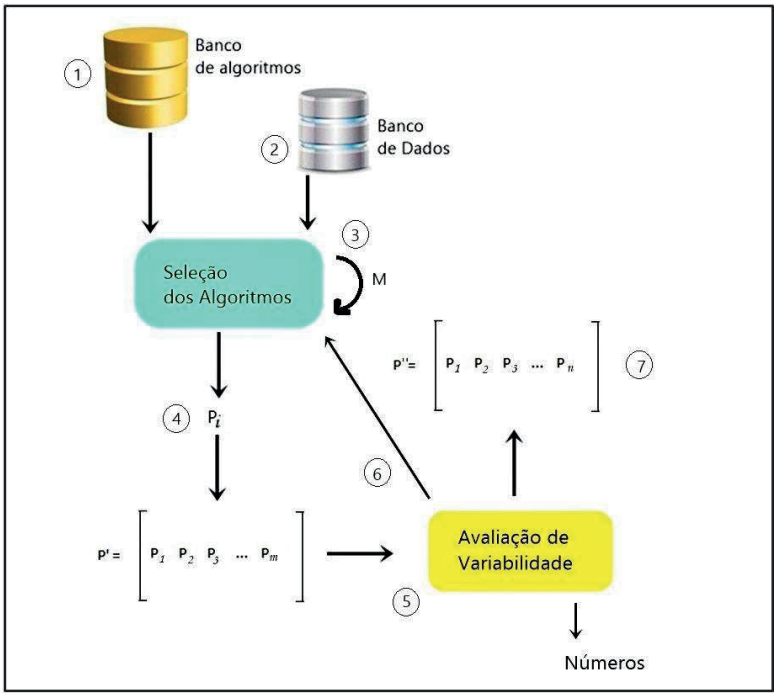

Figura 1. Visão de alto nível das etapas do processo de ensemble clustering.

Fazem-se necessários dois bancos de dados, um de algoritmos de clustering, conforme etapa 1, e outro de dados, conforme etapa 2. Os dados saõ processados pela base de algoritmos, de acordo com etapa 3, e após M iterações serão geradas M partições P_i. Cada partição gerada, conforme etapa 4, será colocada em uma matriz P'. A matriz P' e avaliada na etapa 5, onde são selecionadas as partições com mais alta variabilidade. Nesta etapa, caso uma partição não tenha uma de boa variabilidade, seleciona-se um novo algoritmo para a geração de uma nova partição para substituí-la. Finalmente na etapa 7 uma matriz P'' é produzida com mais alta variabilidade.

A escolha dos algoritmos na etapa 3 é dada aleatoriamente entre algoritmos de clustering dos tipos: algoritmos algomerativos, algoritmos de densidade e algoritmos hierárquicos. Caso o algoritmo selecionado não garanta uma partição com uma boa variabilidade, o algoritmo prevê a troca do algoritmo para geração de uma nova partição.

A permuta deste algoritmo pode ser por outro algoritmo do mesmo conjunto, ou pode ser trocado por um algoritmo dos outros conjuntos.

2 | CONTRIBUIÇÃO DO TRABALHO

O desenvolvimento de um algoritmo de seleção de partições que priorize aquelas que apresentam maior variabilidade entre si resolve o problema de bias introduzido pela pré-existência de partições similares no ensemble.

Estes métodos de agrupamento de dados via combinação de partições, encontram aplicações nos mais diversos segmentos da ciência. Métodos confiáveis de seleção de conjunto de partições diminuirão problemas de bias. Em estatística o conceito de bias está associado à diferença entre a média dos resultados e o valor verdadeiro. Ou seja, é a diferença entre o valor esperado e o valor produzido pelo estimador.

Técnicas de agrupamento de dados visam combinar partições geradas a partir de diversos algoritmos com o intuito de eliminar a necessidade de um estudo prévio acerca dos dados a serem agrupados de modo a definir qual o algoritmo de agrupamento mais adequado. Embora esta seja uma área ativa e recente em reconhecimento de padrões pouco se sabe acerca do impacto da variabilidade do ensemble no resultado final. Acredita-se que este trabalho se configurará como uma contribuição válida de modo a se entender melhor as restrições e áreas de aplicação da metodologia de agrupamento de dados via combinação de partições.

3 | ANÁLISE DE RESULTADOS

Atualmente, as etapas de pesquisa bibliográfica e modelagem do sistema foram finalizadas. Os algoritmos que comporão a base de algoritmos já foram pré- selecionados, foram prospectados conjuntos de dados que comporão a base de dados de testes, e, também desenvolvido os algoritmos para compor a matriz de partições e a analisar a variabilidade das partições.

A etapa em desenvolvimento compõe a matriz de partições escolhendo M algoritmos aleatoriamente e executa uma métrica para análise da variabilidade entre as partições. Caso o algoritmo selecionado não garanta uma partição com uma boa variabilidade, de acordo com a métrica escolhida, o algoritmo é então trocado para geração de uma nova partição.

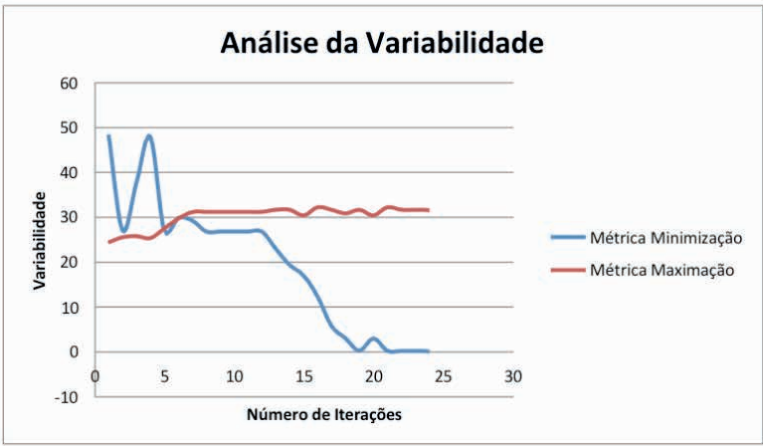

Gráfico 1: Resultado da execução de duas métricas diferentes para análise da variabilidade.

As métricas de variabilidade podem maximizar ou minimizar a variabilidade dependendo do método escolhido para avaliação. Cada métrica avalia uma partição com todas as partições pertencentes a matriz e tem como resultado números. Os resultados dessas aplicações podem ser analisados de duas maneiras como maximização e minimização. O intuito da execução de uma métrica de maximização é aumentar a variabilidade das partições, já à execução de uma métrica de minimização é diminuir a variabilidade das partições.

Com os resultados preliminares demonstrados no gráfico 1, chegamos a conclusão preliminar que a execução do algoritmo maximiza/minimiza a variabilidade intra- partições. O próximo passo para compor os resultados finais, é avaliação do método proposto, ou seja, dado os ensembles com alta variabilidade, serão aplicados métodos de ensemble clustering para verificar se eles realmente geram resultados superiores aos obtidos por métodos ingênuos de geração de partições.

4 | TRABALHOS RELACIONADOS

4.1 A Mixture Model for Clustering

Inúmeros algoritmos de agrupamento são capazes de produzir partições diferentes dos mesmos dados que capturam vários aspectos distintos dos dados [TOPCHY; PUNCH,2004]. O foco nesse artigo é a pesquisa em ensemble clustering, buscando uma combinação de múltiplas partições que proporcionam maior agrupamento geral dos dados fornecidos. No mesmo, relata sobre a maior dificuldade em encontrar uma partição consenso das partições de saída dos vários algoritmos de agrupamento. Outra questão difícil é a escolha do algoritmo de agrupamento para o conjunto.

4.2 Combining Multiple Clusterings Using Evidence Accumulation

O objetivo do agrupamento é particionar um conjunto de objetos não rotulados em grupos homogêneos ou clusters [FRED; JAIN, 2005]. Nesse artigo é relatada a existência de centenas de algoritmos de agrupamento, e a produção de resultados distintos até para o mesmo algoritmo. A abordagem proposta é o conceito de evidência de acumulação de clustering, que mapeia as partições de dados individuais em um conjunto de cluster em uma nova medida de similaridade entre os padrões.

REFERÊNCIAS

Jain M.N. Murty, P. F. A. Data clustering: A review ,1999.

Abdala, D. D. Ensemble and constrained clustering with applications. 2010. TOPCHY, A. K. J. A.; PUNCH, W. A mixture model for clustering ensembles. 2004.

FRED,A. L.; JAIN, A. K. Combining multiple clusterings using evidence accumulation. 2005.