O MÉTODO DE AGRUPAMENTO FUZZY C-MEANS

Henzo Ruan Longo de Souza (PIC), Carina Moreira Costa (Orientador). E-mail: cmcosta@uem.br

Universidade Estadual de Maringá, Centro de Ciências Exatas, Maringá, PR.

Área e subárea do conhecimento: Matemática/Matemática Aplicada

Palavras-chave: Análise de dados; Agrupamento fuzzy; Métodos de otimização;

RESUMO

No âmbito da análise de dados, uma das tarefas mais conhecidas é a de *clustering* ou agrupamento, em que busca-se, a partir de alguma ou várias métricas, dividir os dados coletados em grupos definidos. Uma das abordagens mais conhecidas para particionar os dados coletados é a tarefa de agrupamento fuzzy, em que o pertencimento de um ponto aos grupos é medido usando graus de pertencimento. Essa tarefa pode ser modelada por meio de um problema de otimização não linear com restrições. O método mais conhecido na literatura para resolver esse problema é o Fuzzy C-Means, que foi o objeto de estudo neste trabalho. Esse algoritmo tem a garantia de fornecer uma solução local para o problema, ao invés de uma solução global. No entanto, a partir dos experimentos numéricos realizados, com vários conjuntos de dados, pudemos concluir que o desempenho do método é satisfatório em termos de qualidade da solução e tempo computacional, em comparação à um método exato que fornece soluções globais (ótimas) mas que é muito caro computacionalmente.

INTRODUÇÃO

Nas tarefas de aprendizado de máquina, é comum o processo contínuo de coleta e análise de dados de diferentes fontes e em diversos contextos. Em aplicações como psicologia, biologia, estatística, no próprio aprendizado de máquina, entre outros (Tan et al., 2018, p. 307), é muito importante e útil a tarefa de agrupar os dados massivos e diversos, coletados com algum tipo de critério a fim de conseguir extrair alguma informação ou realizar algum estudo.

A tarefa de agrupamento consiste em dividir um conjunto de dados em grupos, de forma que objetos em um mesmo grupo tenham características similares e que os grupos formados sejam significativos no contexto da aplicação. A atribuição de um ponto à um grupo pode ser definida de duas formas: *hard clustering*, em que cada ponto de um conjunto de dados pertence à somente um grupo e torna-se não pertencente aos grupos restantes; e *fuzzy clustering*, em que o pertencimento de um ponto aos grupos é medido usando-se graus de pertencimento ao invés de atribuições binárias como no caso anterior.

A abordagem do *fuzzy clustering* é especialmente empregada em aplicações em que uma partição binária não faria sentido como, por exemplo, na segmentação do perfil de clientes. A modelagem matemática dessa tarefa pode ser feita por meio de um problema de otimização, o qual envolve expressões não lineares e não convexas, conforme será mostrado a seguir, o que torna a resolução ótima do problema difícil e custosa. Assim, o método mais utilizado para resolver esse problema é um método de busca local, que não tem a garantia de fornecer a solução ótima. Então, o objetivo deste trabalho é avaliar a qualidade das soluções encontradas por este método em comparação à um *solver* que encontra soluções exatas.

REVISÃO DE LITERATURA

O problema de *fuzzy clustering* (Bezdek, 1981, p. 65) pode ser modelado matematicamente da seguinte forma:

$$\begin{aligned} \min_{u,c} \quad & \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m}\|x_j - c_i\|^2 \\ \text{s. a} \quad & \sum_{i=1}^{k} u_{ij} = 1 \quad \forall j = 1,...,n \\ & u_{ij} \in [0,1] \quad \forall j = 1,...,n, \quad i = 1,...,k. \end{aligned}$$

As variáveis de decisão são u e c, em que u é o grau de pertinência de cada ponto aos k grupos, e as variáveis c são os centroides (ou representantes de cada grupo). Os n pontos do conjunto de dados são representados por x, e m é o grau de difusão ou *fuzziness.* Com relação às restrições, as variáveis u determinam o quanto cada um dos pontos pertence a um dado grupo e, portanto, devem assumir valores no intervalo [0,1]. Além disso, para um ponto fixado, a soma dos graus de pertencimento desse ponto aos grupos deve ser igual a 1. Na função objetivo, buscamos minimizar a soma das distâncias ao quadrado, entre os pontos e os centroides.

Para resolver esse problema de otimização, o método mais utilizado na literatura é o algoritmo *Fuzzy C-means* (FCM), que consiste em resolver o problema de forma alternada: primeiro na direção das variáveis u, mantendo os valores de c fixos; e depois na direção das variáveis c, mantendo u fixo, repetindo esse processo até que os centroides não mudem mais ou mudem bem pouco.

Considerando o problema na direção de c, ou seja, com as variáveis u fixadas, temos um problema de otimização irrestrita e, portanto, a solução pode ser calculada como segue:

$$c_i = \frac{\sum_{j=1}^{n} (u_{ij}^m x_j)}{\sum_{j=1}^{n} u_{ij}^m}$$

Logo após, considerando o problema nas variáveis u, o cálculo do grau de pertinência envolve a distância entre os pontos e os centroides, como a seguir:

$$u_{ij} = \left(\sum_{l=1}^{k} \left(\frac{\|x_j - c_i\|^2}{\|x_j - c_l\|^2} \right)^{\frac{1}{m-1}} \right)^{-1}$$

Ambas as expressões são obtidas por condições de otimalidade. A primeira para c, é a condição de primeira ordem, a qual diz que o gradiente da função na solução ótima tem de ser zero. Enquanto que a segunda, para u, é obtida pelas condições KKT (Karush-Kuhn-Tucker). Neste trabalho, realizamos o estudo dessas condições, bem como as demonstrações desses resultados.

A aplicação desse algoritmo, que trabalha de forma alternada entre os dois subproblemas, possui a garantia teórica de encontrar uma solução local, que não é, necessariamente, ótima. Assim, a ideia deste trabalho é verificar a qualidade dessas soluções locais em comparação com as soluções ótimas. Para resolver o problema de forma exata, utilizamos um *solver* de otimização chamado *Gurobi*, que busca a solução global, ou seja, a melhor solução possível para o problema. O desafio se encontra justamente no custo computacional dessa última abordagem, que pode levar horas para encontrar a melhor solução possível. Por outro lado, o algortimo FCM que acabamos de apresentar é barato computacionalmente.

Assim, foram realizados experimentos numéricos utilizando-se diferentes conjuntos de dados, e o problema foi resolvido pelo *Gurobi* e pelo FCM, visando avaliar se as soluções obtidas por cada um eram significativamente diferentes ou não.

RESULTADOS E DISCUSSÃO

Na Tabela 1, apresentamos os resultados comparativos entre o método FCM e o *solver Gurobi*. Na primeira coluna, estão os conjuntos de dados, seguidos pelas dimensões nas três colunas seguintes, em que “n” representa a quantidade de pontos, “d” é a dimensão de cada ponto e “k” é a quantidade de grupos. Na coluna “Obj. Fc”, apresentamos o valor da função objetivo obtido com o método FCM. Na coluna “Obj. Gur”, é mostrado o valor da função objetivo que o Gurobi forneceu dentro de um tempo limite predeterminado. Finalmente, as duas últimas colunas

mostram os valores da métrica Fuzzy Silhouette, que avalia a qualidade do agrupamento.

Tabela 1 – Comparação dos resultados entre Fuzzy C-Means e Gurobi

Conjunto	n	d	k	Obj. Fc	Obj. Gur	Sil. Fc	Sil. Gur
butterfly	15	2	2	26,3282	26,3282	0,58	0,58
twonoisy	97	2	2	170,2129	170,2129	0,62	0,62
iris2d	150	2	3	24,2937	24,2936	0,73	0,73
iris	150	4	3	60,5058	60,5057	0,62	0,62
unbalance	6500	2	8	18,7701	3,5731	0,69	0,83

Como podemos observar, para a maioria dos conjuntos testados, os resultados do FCM foram satisfatórios e muito próximos das soluções globais. Além disso, o FCM geralmente encontrou uma solução local em poucos segundos, enquanto que o *Gurobi* levou horas para determinar a solução exata. Isso justifica e responde à nossa indagação inicial sobre o motivo pelo qual o método FCM é amplamente utilizado na literatura em detrimento de métodos exatos.

CONCLUSÕES

Conclui-se então, que para os conjuntos de dados testados, os resultados do FCM foram satisfatórios e muito próximos das soluções globais encontradas pelo *Gurobi*, justificando o seu uso em comparação à métodos exatos que possuem um custo computacional elevado.

AGRADECIMENTOS

Agradeço a Universidade Estadual de Maringá pela infraestrutura, ao projeto PET pela oportunidade de apresentar os seminários, a minha orientadora e aos meus professores.

REFERÊNCIAS

BEZDEK, J. **Pattern recognition with fuzzy objective function algorithms**. New York: Plenum Press, 1981.

TAN, P.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. **Introduction to data mining**. 2nd ed. London: Pearson, 2018.