Favorecendo o Desempenho do k-Means via Métodos de Inicialização de Centroides de Grupos

Anderson Francisco de Oliveira, Maria do Carmo Nicoletti

Faculdade de Campo Limpo Paulista - FACCAMP
Campo Limpo Paulista - SP, Brasil

{anderson@asmec.br,carmo@cc.faccamp.br}

Abstract. *Clustering can be stated in a simplistic way as: given a set of patterns X, find the best way to divide them into disjoint groups of patterns, so that their union restores X. The simplest category of clustering algorithms is the partitional. Partitional algorithms organize the data patterns in a clustering of disjoint clusters. The k-Means algorithm is one of the most popular and well-known partitional algorithm. However, its initialization step can have a negative impact on the produced clustering. This project investigates a few initialization methods proposed in the literature, aiming at improving the performance of k-Means algorithm by using a more efficient initialization. Also, as another goal, the convergence speed of the k-Means, when using each of the initialization methods considered, will be investigated.*

Resumo. *Agrupamento pode ser estabelecido de uma maneira simplista como: dado um conjunto de padrões X, encontrar a melhor forma de dividi-los em grupos disjuntos de padrões, de maneira que a união de tais grupos recomponha X. A categoria mais simples de algoritmos de agrupamento é a particional. Algoritmos particionais organizam os padrões de dados de um conjunto em vários grupos disjuntos. O algoritmo k-Means é um dos mais conhecidos dentre os métodos particionais. No entanto, a sua inicialização pode impactar negativamente o agrupamento que produz. Este projeto investiga alguns métodos de inicialização propostos na literatura, com o objetivo de melhorar o desempenho do algoritmo k-Means, por meio do uso de uma inicialização mais eficiente. Também será investigada a velocidade de convergência do k-Means, quando do uso de cada um dos métodos de inicialização investigados.*

1. Introdução e Contextualização

A área de Aprendizado de Máquina (AM) investe, entre outros, no estudo e pesquisa de estratégias, métodos e algoritmos que permitem que computadores possam aprender. Os muitos algoritmos já propostos e utilizados em AM se dividem, a grosso modo, em dois grandes grupos: os de aprendizado supervisionado e os de aprendizado não-supervisionado. As referências [Jain & Dubes 1988] [Mitchell 1997] [Duda *et al*. 2001] [Theodoridis & Koutroumbas 1999] [Jain 2010] [Witten *et al.* 2011] apresentam revisões de muitos desses algoritmos. Para que o aprendizado automático seja viável, seja ele supervisionado ou não supervisionado, é mandatório a existência de um conjunto de dados, a partir do qual o aprendizado acontece (geralmente como um processo de generalização de tal conjunto).

Via de regra os dados são um conjunto de padrões, cada uma deles descrito como um vetor de valores de atributo e, eventualmente, uma classe associada. A classe identi-

fica o conceito (ou categoria) à qual cada padrão, descrito por determinados valores de atributo, pertence. Os chamados algoritmos supervisionados fazem uso da informação da classe do padrão, para aprender. Nem sempre, entretanto, padrões de conjuntos de dados disponibilizados têm uma classe associada. Para o aprendizado automático baseado em padrões não têm uma classe associada, são usados algoritmos que podem ser caracterizados como "agrupadores de dados", uma vez que buscam organizar os dados em um conjunto de grupos, de acordo com critérios variados. Tais algoritmos são identificados pelo nome geral de algoritmos de agrupamento e são os representantes mais usados dos chamados algoritmos não supervisionados. Na literatura pode ser encontrado um número substancial de diferentes algoritmos de agrupamento, bem como um número razoável de taxonomias que tentam organizar tais algoritmos, de acordo com algumas de suas características básicas (ver, por exemplo [Theodoridis & Koutrumbas 1991]).

De interesse neste trabalho de pesquisa é um dos mais conhecidos algoritmos de agrupamento chamado k-Means [MacQueen 1967] que, desde a sua proposta em 1967, continua sendo usado em uma grande diversidade de domínios de dados, devido à sua simplicidade, fácil implementação e rapidez em execução. O k-Means é caracterizado como um algoritmo particional que, dado um conjunto de padrões como entrada, tem por objetivo encontrar uma partição do conjunto em k grupos disjuntos. Como já especificado no próprio nome do algoritmo, o k é também um parâmetro de entrada para o algoritmo, fornecido pelo usuário, e que representa o número de grupos que o agrupamento, a ser induzido pelo algoritmo, deve ter. O k-Means inicia a construção do agrupamento por meio da escolha randômica dos centroides dos k grupos (de padrões) a serem construídos.

Para um dado conjunto de padrões e, devido ao fato da escolha inicial dos k centroides de grupos ser randômica, o k-Means nem sempre induz o mesmo agrupamento, em duas execuções distintas do mesmo algoritmo, com o mesmo conjunto de dados e o mesmo valor para o parâmetro k; esse fato pode ser um problema em certos domínios de dados. Na literatura podem ser encontrados vários métodos que buscam resolver o problema da inicialização dos centroides de grupos, bem como alguns trabalhos que fazem revisões de alguns desses métodos, tais como aquelas apresentadas em [Peña *et al.* 1999] [Khan & Ahmad 2004] [Celebi *et al.* 2013]. Este projeto de pesquisa está voltado à investigação de alguns desses métodos, com o objetivo de identificar reais contribuições ao problema e avaliar quão factíveis e representativos efetivamente são. O projeto contempla a investigação da velocidade de convergência do k-Means, quando do uso de cada um dos métodos de inicialização a serem investigados. O projeto prevê, quando de sua finalização, disponibilizar um sistema computacional com as implementações do k-Means e dos vários métodos de inicialização de centroides de grupos que serão investigados, com vistas à experimentação, aprendizado e ensino.

2. O Algoritmo k-Means

Como resumidamente apresentado em [Witten *et al.* 2011], tendo como entrada um conjunto de N padrões (ou pontos) CP = $\{p_1, p_2, ..., p_N\}$ e um valor (inteiro) atribuído ao parâmetro k, o algoritmo k-Means inicia escolhendo, randomicamente, k padrões, que representam k centroides de grupos (centroide é caracterizado como a média dos padrões associados a um grupo). Cada padrão de CP, então, é atribuído ao grupo cujo centroide lhe seja mais próximo, por meio do cálculo da distância (euclidiana, geralmente) de cada padrão, a cada um dos k centroides de grupos considerados. A seguir, a média

dos padrões atribuídos a cada grupo (isto é, os respectivos centroides de grupos) é calculada. Esses centroides passam, então, a ser os novos centroides de grupos e todo o processo é repetido, com os novos centroides de grupos. O processo iterativo continua até que os mesmos padrões sejam atribuídos aos mesmos grupos, em iterações consecutivas, um indicativo que os centroides de grupos atingiram estabilidade e assim permanecerão. Uma vez que o processo iterativo tenha se estabilizado, cada padrão é atribuído ao grupo associado ao seu centroide de grupo mais próximo, processo que pode ser matematicamente parafraseado como tendo efeito de minimizar o total dos quadrados das distâncias de todos os padrões aos seus respectivos centroides de grupos. Esse mínimo, entretanto, é local e não existe garantia que seja um mínimo global. Os grupos resultantes de um agrupamento induzido pelo k-Means são tão sensíveis à escolha inicial dos centroides de grupos que uma pequena mudança no conjunto dos centroides de grupos escolhidos inicialmente, pode implicar a criação de um agrupamento completamente diferente. Para a obtenção de bons resultados com o k-Means, usualmente o que se faz na prática é executá-lo um determinado número de vezes e, a cada vez, com um conjunto diferente de centroides de grupos.

Como apontado em [Han *et al.* 2012], a complexidade em tempo do k-Means é dada por O(Nkt), em que N é o número total de padrões, k é o número de grupos e t é o número de iterações. Normalmente k << N e t << N, o que torna o algoritmo relativamente escalável e eficiente, quando do processamento de um grande volume de dados. Na literatura podem ser encontradas várias variações do k-Means original e, geralmente, tais variações diferem com relação à seleção inicial dos centroides de grupos, cálculo da dissimilaridade (distância) e estratégias para o cálculo dos centroides. A Figura 1 apresenta um pseudocódigo simplificado do algoritmo k-Means, inspirado naquele encontrado em [Han *et al.* 2012].

procedure *k-Means(CP,k,AG)*
Input: CP = {p_1, p_2, ..., p_N} %conjunto de padrões de dados a ser agrupado
k % número de grupos a ser criado
Output: {G_1,G_2,...G_k} %agrupamento formado por k grupos de padrões de dados
begin
(1) escolher arbitrariamente k padrões ∈ CP, como centroides dos grupos G_1,G_2,...G_k respectivamente % nesse passo cada grupo é definido apenas pelo centroide
(2) **repeat**
(3) (re)atribuir cada padrão p_i ∈ CP ao grupo associado ao centroide que lhe seja mais próximo;
(4) atualizar os centroides de cada um dos k grupos, como a média dos padrões a ele associados;
(5) **until** nenhuma alteração aconteça.
end.
return AG = {G_1,G_2,...G_k}
end_procedure

Figura 1. Pseudocódigo simplificado do k-Means.

3. Métodos de Inicialização Considerados

Como comentado anteriormente, na literatura podem ser encontrados inúmeros métodos que se propõem a sanar a deficiência do k-Means, com relação à sua proposta original de, no seu primeiro passo, selecionar randomicamente k padrões do conjunto de

dados e promove-los a centroides de grupos. Dentre os até agora pesquisados, os escolhidos para a continuação do trabalho e aprofundamento em seus detalhes técnicos são:

(1) Em [Duda *et al.* 2001] é proposto um método recursivo para a inicialização dos centros de grupos, por meio da execução de k problemas de agrupamento. Uma variação deste método consiste em considerar todo o conjunto inicial de padrões e perturbá-lo randomicamente k vezes.

(2) Os autores em [Jain & Dubes 1988] usaram o k-Means um grande número de vezes, com seleção randômica dos centros de grupos e, então, selecionaram a média dos centros de grupos obtidos como o conjunto inicial de centros de grupos.

(3) Em [Bradley & Fayyad 1998] é proposto um algoritmo de refinamento que constrói um conjunto de pequenas sub-amostras do conjunto original de padrões e, então, realiza um processo de agrupamento, usando o k-Means, em cada uma delas. Todos os centroides de todas as sub-amostras são, então, agrupados, pelo k-Means, usando os k centroides de cada sub-amostra como os centroides iniciais. Os centroides do agrupamento final que produzir o menor erro de agrupamento são, então, usados como centroides iniciais, para o agrupamento do conjunto original de padrões, usando o k-Means.

(4) O algoritmo CCIA (*Cluster Center Initialization Algorithm*) [Khan & Ahmad 2004] foi baseado em duas observações associadas a processos de agrupamento: (1) alguns padrões são muito semelhantes entre si e, devido a isso, eles pertencem ao mesmo grupo, independentemente da escolha inicial dos centros de grupos; (2) um atributo (dentre os que descrevem os padrões) pode fornecer informação a respeito dos centros iniciais de grupos. O algoritmo é voltado para dados descritos por atributos contínuos.

(5) A proposta restrita a agrupamentos com dois grupos, descrita em [Li 2011], é baseada no algoritmo NN [Cover & Hart 1967]. Esta abordagem implementa o algoritmo CIT que usa os conceitos de vizinho mais próximo entre dois padrões, vizinhos mais próximos entre dois pares de padrões e dissimilaridade em pares de vizinhos mais próximos e assume como válidas quatro suposições teóricas.

(6) O algoritmo descrito em [Erisoglu *et al.* 2011] é baseado na seleção dos dois principais atributos que descrevem os padrões de dados. Os atributos são selecionados de acordo com o coeficiente máximo de variação e o valor mínimo absoluto e correlação e o algoritmo continua, abordando o conjunto original descrito apenas pelos dois atributos selecionados.

A metodologia para o desenvolvimento do projeto prevê: (1) estudo e entendimento em detalhes, de cada algoritmo; (2) implementação dos algoritmos; (3) escolha e criação de conjuntos de dados que reflitam situações corriqueiras bem como situações limites (particularmente, o projeto contempla o uso de conjuntos de dados que já tenham sido utilizados para evidenciar o potencial de alguns dos algoritmos escolhidos, cujos resultados tenham sido publicados); (4) seleção de índices de validação, também a serem implementados; (5) parte experimental, consistindo no uso de cada um dos algoritmos de inicialização acoplado ao k-Means (destituído de sua inicialização randômica), para a avaliação dos agrupamentos induzidos.

Referências

[Bradley & Fayyad 1998] Bradley, P. S.; Fayyad, U. (1998) Refining initial points for k-means clustering, in: Proc. of the 15th International Conference on Machine Lear ing, pp. 91-99.

[Celebi *et al.* 2013] Celebi, M. E.; Kingravi, H. A.; Vela, P. A. (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm, Expert Systems with Applications, v. 40, pp. 200-120.

[Cover & Hart 1967] Cover, T.; Hart, P. (1967) Nearest neighbor pattern classification, IEEE Transactions on Information Theory, v. IT 13, pp. 21-27.

[Duda *et al.* 2001] Duda, R. O.; Hart, P. F.; Stork, D. G. (2001) Pattern Classification, USA: John Wiley & Sons, Inc.

[Erisoglu *et al.* 2011] Erisoglu, M.; Calis, N.; Sakalliouglu, S. (2011) A new algorithm for initial cluster centers in k-means algorithm, Pattern Recognition Letters, v. 32, pp. 1701-1705.

[Han *et al.* 2012] Han, J.; Kamber, M.; Pei, J. (2012) Data Mining Concepts and Techniques, 3rd. Ed., Amsterdam: Morgan Kaufmann Publishers.

[Jain & Dubes 1988] Jain, A.K., Dubes, R.C. (1988) Algorithms for Clustering Data, Prentice Hall.

[Jain 2010] Jain, A.K. (2010) Data clustering: 50 years beyond k-Means, Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666.

[Khan & Ahmad 2004] Khan, S. S.; Ahmad, A. (2004) Cluster center initialization algorithm for k-Means clustering, Pattern Recognition Letters, v. 25, pp. 1293-1302.

[Li 2011] Li, C. S. (2011) Cluster center initialization method for k-Means algorithm over data sets with two clusters, Procedia Engineering, v. 24, pp. 324-328.

[MacQueen 1967] MacQueen, J. B. (1967) Some methods for classification and analysis of multivariate observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

[Maedeh & Suresh 2013] Maedeh, A.; Suresh, K. (2013) Design of efficient k-means clustering algorithm with improved initial centroids, MR International Journal of Engineering and Technology, v. 5, no. 1, pp. 33-37.

[Mitchell 1997] Mitchell, T. M (1997) Machine Learning, USA: McGraw-Hill.

[Peña *et al.* 1999] Peña, J.M. ; Lozano, J.A. ; Larrañaga, P. (1999) An empirical comparison of four initialization methods for the K-Means algorithm, Pattern Recognition Letters, vol. 20, pp. 1027-1040.

[Theodoridis & Koutroumbas 1999] Theodoridis, S.; Koutroumbas, K. (1999) Pattern Recognition, USA: Academic Press.

[Witten *at al.* 2011] Witten, I. H.; Frank E.; Hall, M. A. (2011) Data Mining: Practical Machine Learning Tools and Techniques, 2nd. Ed., Amsterdam: Morgan Kaufmann Publishers.