Comparação entre os algoritmos *K-Means* e *Dynamic Cluster* em imagens digitais

Arthur Scardini Domingues[1], Maria Luzia Silva de Carvalho[1], José Washington Vidal Morais Neto[1], Angélica Félix de Castro[1]

[1]Programa de Pós-Graduação em Ciência da Computação UFERSA/UERN – Universidade Federal Rural do Semi-Árido (UFERSA)
59526-900 - Mossoró – RN – Brasil

arthurdomingues91@hotmail.com, luziacc2009@gmail.com, moraisneto_@hotmail.com, angelica@ufersa.edu.br

Abstract. *The present work aimed to compare two clustering algorithms (K-Means and Dynamic Cluster) in digital images. These were submitted to the same tests in order to evaluate their performance regarding the quality of the response. The results showed the superiority of the Dynamic Cluster algorithm in relation to K-Means.*

Resumo. *O presente trabalho, teve como objetivo realizar o comparativo entre dois algoritmos de clusterização (K-Means e Dynamic Cluster) em imagens digitais. Estes foram submetidos aos mesmos testes a fim de avaliar sua performance quanto a qualidade da resposta. Os resultados mostraram a superioridade do algoritmo Dynamic Cluster em relação ao K-Means.*

1. Introdução

Linden (2015) define análise de agrupamento, ou *clustering*, como o nome dado para o grupo de técnicas computacionais cujo propósito consiste em separar objetos em grupos, baseando-se nas características que estes objetos possuem. A ideia básica objetiva colocar em um mesmo grupo objetos que sejam similares de acordo com algum critério pré-determinado.

Os algoritmos de agrupamento particionam um conjunto de objetos em agrupamentos (Manning e Schutze, 2003). Normalmente, objetos são descritos e agrupados por um especialista usando um conjunto de atributos e valores, não existindo nenhuma informação sobre a classe ou categoria dos objetos. O objetivo dos algoritmos de agrupamento é colocar os objetos similares em um mesmo grupo e objetos não similares em grupos diferentes.

Existem dois tipos de estruturas produzidas por algoritmos de agrupamento: (i) não hierárquicos ou planos; e (ii) agrupamentos hierárquicos (Barth, 2013). Os agrupamentos planos simplesmente contêm um certo número de agrupamentos e a relação entre os agrupamentos é geralmente não-determinada. A maioria dos algoritmos que produzem agrupamentos planos são iterativos: eles iniciam com um conjunto inicial de agrupamentos e realocam os objetos em cada agrupamento de maneira iterativa, até uma determinada condição de parada. Um agrupamento hierárquico é representado por

uma árvore, onde os nós folhas são os objetos e cada nó intermediário representa o agrupamento que contêm todos os objetos de seus descendentes (Mitchel, 1997).

O objetivo desse trabalho consiste em realizar uma análise comparativa entre os algoritmos de clusterização *K-Means* e *Dynamic Cluster* em imagens digitais a fim de analisar a performance e qualidade obtidas. Ambos foram selecionados pelos seguintes motivos: o *K-Means* é um algoritmo conhecido e bem aceito pela academia, enquanto que o *Dynamic Cluster* é menos conhecido e com uma complexidade maior. O motivo de aplica-los em imagens digitais foi a característica matemática delas, usando o conceito de matrizes e pixels.

Esse artigo está organizado da seguinte maneira: a seção 2 apresenta uma revisão da literatura, com uma explicação mais detalhada sobre mineração de dados, clusterização e os algoritmos *K-Means* e *Dynamic Cluster*; na seção 3 foram exibidos alguns trabalhos relacionados; na seção 4 foi realizada a comparação entre os dois algoritmos envolvidos e; por fim, a seção 5 apresenta os resultados e discussão desse trabalho.

2. Revisão da Literatura

Abaixo, alguns conceitos serão explicados de forma a facilitar a compreensão do presente trabalho.

2.1. Mineração de Dados

O termo mineração de dados se refere ao uso de um agregado de métodos e tecnologias que permitem automatizar a busca em grandes volumes de dados por padrões e tendências que não são detectáveis por análises mais simples (Fayyad, 1997). Elmasri e Navathe (2005) reafirmam a ideia de que mineração de dados se refere à descoberta de novas informações em função de padrões em grandes quantidades de dados. É possível por meio da mineração de dados encontrar respostas para questões que por técnicas mais simples, não seriam possíveis.

Witten *et al* (2005) apresentam algumas das áreas nas quais a mineração de dados é aplicada de forma satisfatória: auxílio em pesquisas biométricas, retenção de clientes, identificação de perfis para determinados produtos, mineração e dados em imagens, entre outros exemplos aplicados.

2.2. Clusterização

Clustering ou clusterização é uma técnica de mineração de dados para realização de agrupamentos de dados (Hartigan, 1975). Segundo JAIN *et al.* (1999), o processo de clusterização é a classificação não-supervisionada de dados, formando agrupamentos ou *clusters*. Representa uma das principais etapas de processos de análise de dados, denominada análise de clusters. A Clusterização é uma técnica bastante útil em diversas soluções de problemas nas mais diversas áreas do conhecimento para o estudo e compreensão do comportamento de dados. De uma forma geral, Clusterização consiste em agrupar elementos de uma determinada base de dados em subgrupos, onde esses elementos que compartilham as mesmas características estarão agrupadas em um mesmo cluster.

2.3. Algoritmo *K-Means*

O *K-Means* é uma técnica de clusterização destinada ao agrupamento de n conjunto de dados (instâncias) em k grupos. Ele é matematicamente capaz de lidar com qualquer modelo de dados. As n instâncias são métricas de distância com os k grupos (também conhecidos por protótipos), passando iterativamente a pertencer ao grupo que estiver mais próximo. Os k grupos também vão sendo reajustados a cada iteração (Nogare, 2016).

De acordo com Santos (2008), a métrica utilizada, o formato do grupo e até a intepretação de um dado pode variar dependendo do formato/tipo do conjunto de entrada. Para os dados em questão (imagens digitais), o conjunto de entrada pode ser definido como janela, que vem a ser a representação de uma unidade para os processos de segmentação. Ela pode ser composta por um ponto, ou uma submatriz da imagem principal. Um pixel possui 8 vizinhos. Com o aumento do tamanho da janela, essa vizinhança sobe proporcionalmente.

Medir a distância entre objetos e moldes significa definir o quão próximos estão os seus marcadores sob a lógica de um sistema de coordenadas. Existem várias métricas para calcular distância: L1, L2, *chebyshev*, *procrutes*. E de acordo com o formato da janela, a distância do protótipo à instância pode ser vista como a mediana, média, moda, entre outros nomes (Santos, 2008).

O *K-Means* utiliza todos esses conceitos, calculando a distância de todo o conjunto, em uma quantidade indefinida de iterações. A cada iteração, os grupos são recalculados e os elementos redistribuídos entre eles. Quando a quantidade de mudanças de grupos entre os elementos for menor do que um limiar (comumente muito menor do que a quantidade de janelas), o algoritmo chega em seu ponto de parada. Assim, o processo de clusterização terá terminado (Nogare, 2016). Um fluxograma de execução do *K-Means* pode ser visto na Figura 1 abaixo:

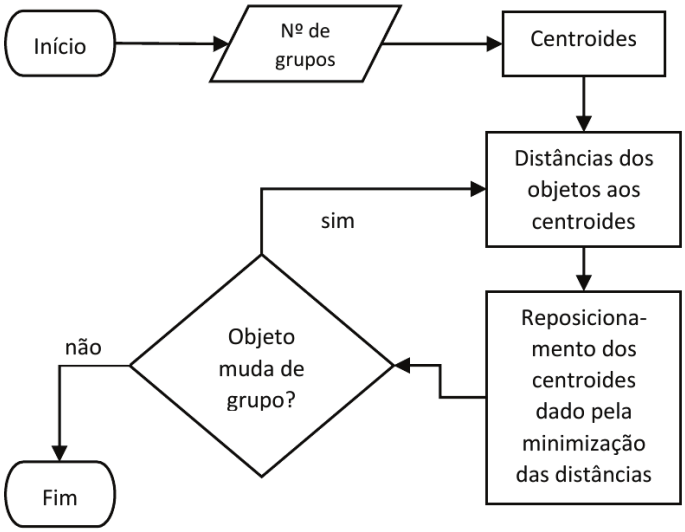


Figura 1. Fluxograma de execução do Algoritmo K-Means

Para o problema da binarização, só existem dois grupos: completamente opaco ou completamente transparente (0 e 1). Esse é um caso bem específico do *K-Means*,

utilizado para descobrir o que é objeto e o que é fundo nas imagens. Apesar de somente terem dois grupos, esse algoritmo possui uma complexidade assintótica alta por dois fatores: primeiramente, cada iteração opera todos os elementos da matriz, além disso, a quantidade de iterações depende de como se comportam as instâncias. Dependendo do tamanho da janela, o algoritmo pode levar horas para ser finalizado. Consiste em uma técnica de segmentação que atribui a cada janela, o valor 0 ou 1; produzindo em sua saída, uma imagem com dois tipos de cores: preto completo (0), ou branco (1). Assim, há uma separação ente os objetos de maior ou menor importância.

2.4. *Dynamic Clustering*

O *Dynamic Clustering* é uma técnica baseada no algoritmo KNN (*K-Nearest Neighbors*), que se destina a agrupar elementos de um conjunto por meio de limiarização, ou seja, dividindo os grupos a partir de um limiar (Shafeeq e Hareesha, 2012). Para entender como acontece a clusterização por meio desse algoritmo é necessário que se tenha em mente o que é um histograma de cores.

Conforme Silva (2007), em processamento de imagens, trabalha-se sempre com os tons de cinza (*Digital Numbers* ou DNs) atribuídos aos pixels de uma imagem. O histograma é uma das formas mais comuns de se representar a distribuição dos DNs de uma imagem, e possivelmente a mais útil em processamento digital de imagens (Crósta, 1993). Ele fornece a informação sobre quantos pixels na imagem possuem cada valor possível de DN (que, no caso das imagens de 8 bits, variam de 0 a 255) ou, de forma equivalente, qual a proporção da imagem que corresponde a cada valor de DN.

O histograma pode demonstrar como estão distribuídas as diferentes intensidades de cores. Também pode ser utilizado para saber se determinada intensidade existe ou não. O primeiro passo para o processamento do *Dynamic Clustering* é descobrir o histograma de cores da imagem (Figura 2): com ele é possível obter a distribuição de intensidades com clareza.

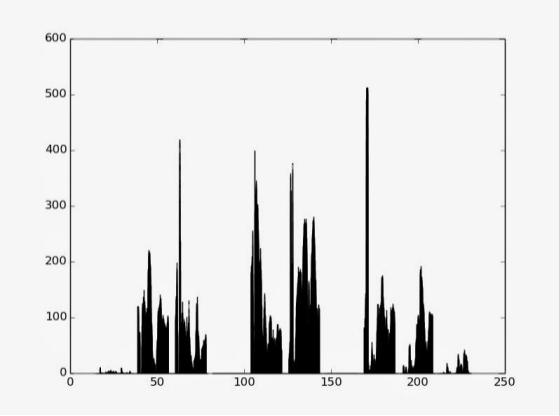


Figura 2. Exemplo de Histograma

O objetivo desta técnica é dinamicamente definir um limiar que agrupe ao elementos da melhor maneira. Uma das maneiras de detectar o limiar é analisar o histograma da imagem. Para o problema da binarização, um só limiar é suficiente pois esse limiar consiste em separar uma imagem em regiões de interesse e não interesse

através da escolha de um ponto de corte. Conforme a Figura 3 abaixo, o limiar T é definido conforme as iterações vão o dispondo mais à direita ou mais à esquerda no histograma.

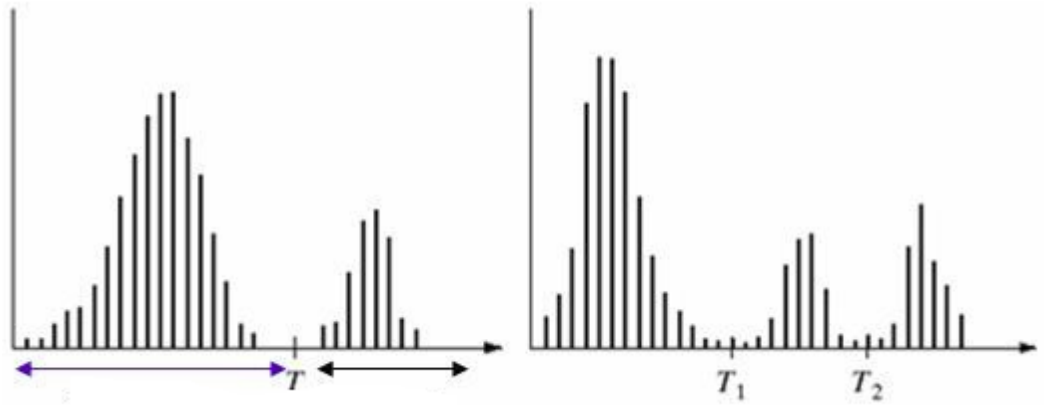


Figura 3. Limiarização

Esse algoritmo tem uma complexidade bem reduzida quando comparado ao *K-Means*; uma vez que em poucas iterações o limiar já atinge uma posição perto da ótima. Sua condição de parada também é dada quando o limiar atinge um valor de alteração quase nulo, assim havendo uma quantidade irrisória de troca de grupos. Neste caso, não há necessidade de indicação em cada iteração de cada elemento a um grupo. Isso só acontece ao final do processamento quando o limiar T tem atingido sua posição de otimalidade.

3. Trabalhos Relacionados

Prass (2004) apresentou um estudo comparativo dos principais modelos de algoritmos de Análise de Agrupamento (*Cluster Analysis*) existentes na literatura e implementados em softwares, visando o seu uso no processo de descoberta de conhecimentos em grandes bancos de dados (*Knowledge Discovery in Databases* - KDD).

Ele realizou um estudo comparativo dos principais algoritmos de Cluster existentes na literatura e implementados na Linguagem R visando o seu uso no processo de descoberta de conhecimentos em grandes bancos de dados. Os algoritmos foram avaliados com dados reais (dados de imóveis da cidade de Belém do Pará) e simulados (duzentos e dez mil registros gerados com o auxílio do software SPSS) de acordo com o seu método de formação (Hierárquico, Partição, Baseado em Modelo, Baseado em Grade e Baseado em Densidade) e também pela medida de distância que expressa a similaridade ou dissimilaridade entre os objetos.

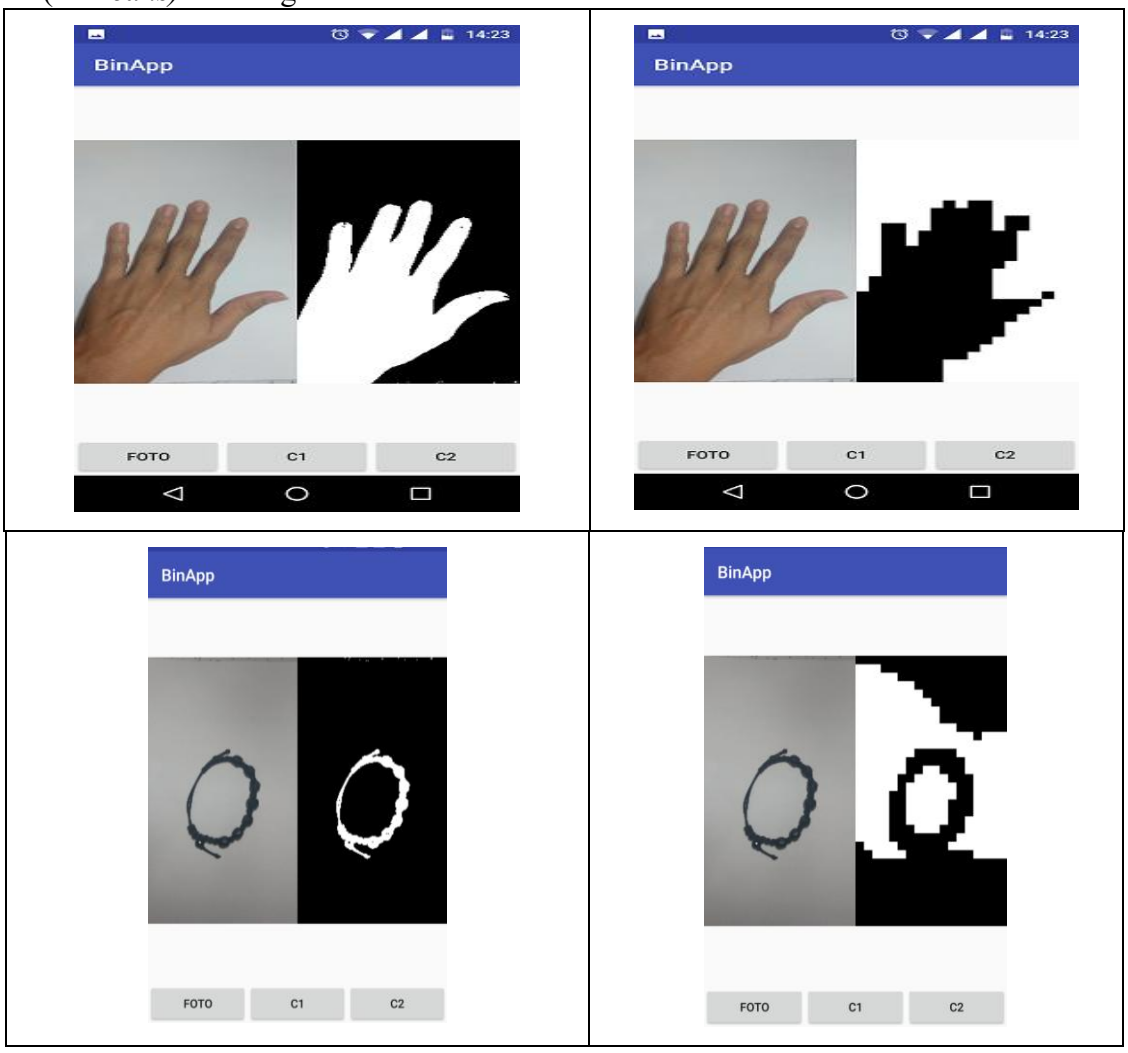
Cargnelutti Filho *et al* (2008) realizou uma comparação de métodos de agrupamento para o estudo da divergência genética em cultivos de feijão. Trevisan *et al* (2013) comparou algoritmos de clustering hierárquico em dados reais, tendo como problemática um caso específico da agricultura: o trabalho apresentou um estudo de caso considerando informações físico-químicas do solo e aspectos de coloração das folhas de soja para a análise exploratória desses dados por meio de aprendizado de máquina não-supervisionado. Foi feita uma comparação entre diferentes algoritmos de

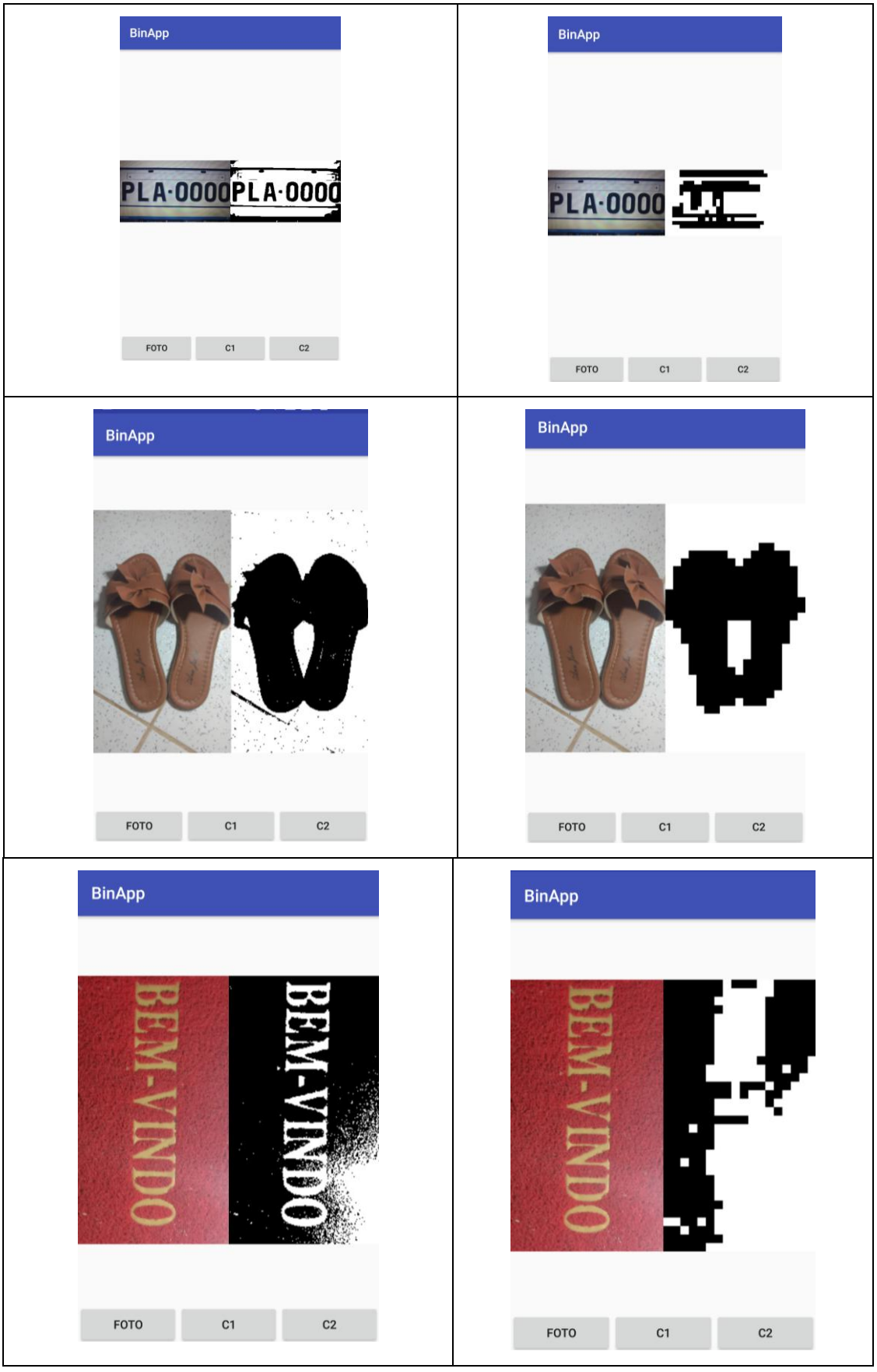
agrupamento hierárquico para identificar os métodos mais adequados para esse domínio de aplicação e encontrar evidências para facilitar a identificação automática das necessidades específicas de cada subárea da lavoura.

4. Comparação realizada entre o uso do Dynamic Clustering e o K-Means

O estudo foi composto por uma comparação entre as duas técnicas: *Dynamic Clustering* (C1) e *K-Means* (C2). Foram utilizadas 30 amostras de imagens e postas em processamento com ambos os algoritmos. Os algoritmos foram implementados em um aplicativo para um dispositivo Android e utilizando a biblioteca Bitmap, nativa do ambiente. Assim, os dois algoritmos foram dispostos em uma situação de igualdade quanto a hardware, software e implementação.

Como visto anteriormente, o C2 depende do tamanho da janela para calcular com mais ou menos velocidade e qualidade. Com a janela de tamanho 1, não foi possível obter resultados. Então foi utilizada uma de tamanho 3, ou seja, com 9 pixels. Alguns testes podem ser vistos nas imagens dispostas na Figura 4 abaixo. Leve em consideração que o C1 (*Dynamic Cluster)* é a imagem mais à esquerda, enquanto que o C2 (*K-Means)* é a imagem mais à direita.




BinApp
PLA-0000
PLA-0000
FOTO
C1
C2
BinApp
PLA-0000
FOTO
C1
C2
BinApp
FOTO
C1
C2
BinApp
FOTO
C1
C2
BinApp
BEM-VINDO
BEM-VINDO
FOTO
C1
C2
BinApp
BEM-VINDO
FOTO
C1
C2

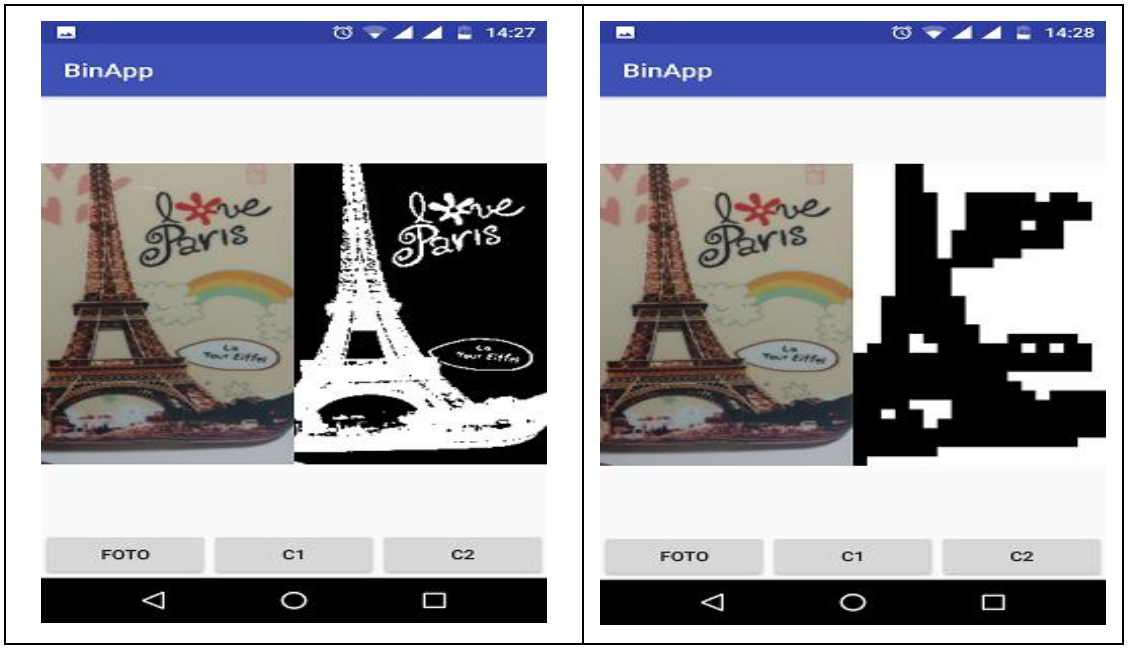


Figura 3. Comparativo entre os algoritmos Dynamic Cluster (C1) e K-Means (C2)

É possível visualizar na Figura 3 que o C1 (lados esquerdos das figuras) se mostrou mais eficiente do que o C2 (lados direitos) em ambos os casos. A visualização das imagens após a compilação dos algoritmos foi bem melhor no uso do C1 – *Dynamic Cluster*.

5. Resultados e Discussão

Os resultados do algoritmo C1 foram superiores aos do C2. Como pôde ser observado nas imagens acima, o algoritmo de limiarização obteve uma binarização favorável em todas imagens, onde a qualidade das imagens foi superior em todos os testes: perceba que todas as imagens à esquerda (aplicação do algoritmo C1) foram melhores do que as imagens à direita (aplicação do algoritmo C2).

O C2 foi posto à prova com suas configurações máximas (janela de 1 pixel) em um celular de 8 núcleos de processamento, mas ainda assim não obteve resultado: houve um estouro de memória, travando e fechando a aplicação inteira. O C1 em todos os casos atingiu um resultado satisfatório, rápido, conseguindo identificar os objetos e o fundo das imagens.

Referências

Barth, F. J. (2013). “Algoritmos de Agrupamento - Aprendizado Não Supervisionado”. Slides disponíveis em: <http://fbarth.net.br/materiais/docs/am/agrupamento.pdf>. Acesso em: 18 mar 2018.

Cargnelutti Filho, A; Ribeiro, N. D.; Reis, R. C. D.; Souza, J. R. e Jost, E. (2008). “Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão”. Ciência Rural, Santa Maria, v.38, n.8, p.2138-2145, nov, 2008. ISSN 0103-8478.

CRÓSTA, A. P. (1993). “Processamento Digital de Imagens de Sensoriamento Remoto”. UNICAMP.

Elmasri, R e Navathe, S. B. (2005). “Sistemas de Banco de Dados 4” ed. São Paulo: Pearson Addison Esley.

Fayyad, U.M. (1997). Editorial. Data Mining and Knowledge Discovery Journal. 1 (1), 5-10.

Hartigan, J.A. (1975). “Clustering Algorithms”, John Wiley.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). “Data clustering: a review”. ACM Computing Surveys, 31(3), 264 – 323.

Linden, R. (2015) “Técnicas de Agrupamento”, Revista de Sistemas de Informação da FSMA, n. 4, pp. 18-36

Manning, C. D. e Schutze, H. (2003). “Foundations of Statistical Natural Language Processing”. MIT Press.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Nogare, D. (2016). “Engenharia do Conhecimento e Sistemas Especialistas”. Disponível em: < http://www.diegonogare.net/2015/08/entendendo-como-funciona-o-algoritmo-de-cluster-k-means/>. Acesso em: 08 mar. 2018.

Prass, F. S. (2004). “Estudo comparativo entre algoritmos de análise de agrupamentos em data mining”. Dissertação de Mestrado. Universidade Federal de Santa Catarina.

Santos, R. (2008). “Conceitos de Mineração de Dados Multimídia”. WebMedia, 2008. Disponível em: <http://www.lac.inpe.br/~rafael.santos/Docs/WebMedia/2008/mmdm.pdf>. Acesso em: 18 mar 2018.

Shafeeq, B. M. A. e Hareesha, K. S. (2012). “Dynamic Clustering of Data with Modified K-Means Algorithm”. 2012 International Conference on Information and Computer Networks (ICICN 2012).

SILVA, A. M. (2007). “Curso Processamento digital de imagens de satélite”. Centro de Eventos da PUCRS - de 07 a 12 de outubro de 2001. Porto Alegre - RS. Disponível em: www.cartografia.org.br. Acesso em: 08 mar. 2018.

Trevisan, T. B.; Ziglioli, M.; Mallmann, A. A.; Metz, J. e Paula Filho, P. (2013). “COMPARAÇÃO DE ALGORITMOS DE CLUSTERING HIERÁRQUICO EM DADOS REAIS: UM ESTUDO DE CASO NA AGRICULTURA”. Revista Eletrônica Científica Inovação e Tecnologia. Disponível em: <https://revistas.utfpr.edu.br/recit/article/viewFile/214/pdf>. Acesso em: 08 mar. 2018.

Witten, I. H; Frank, E. (2005) “Data Mining - Practical Machine Learning Tools and Techniques”, Elsevier.