Aplicação de Algoritmos de Clusterização em uma Base de Dados de Reservas de Hotéis

Application of Clustering Algorithms in Hotels Reservation Datasets

Pedro Alexandre de Araújo Aguiar[1] orcid.org/0000-0002-9973-763X
Clodomir Joaquim de Santana Junior[1] orcid.org/0000-0001-7869-7184
Carmelo José Albanez Bastos Filho[1] orcid.org/0000-0002-0924-5341

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Pernambuco, Brasil.

E-mail do autor principal: paaa@ecomp.poli.br

Resumo

Este artigo faz uma análise da aplicação dos algoritmos de clusterização K-Means e Fuzzy C-Means. O estudo de caso visa identificar perfis de clientes de uma agência de viagens online, com o objetivo de melhorar a eficácia do envio de ofertas através de e-mail marketing, possibilitando o envio de anúncios personalizados para cada perfil. O processo de clusterização foi feito baseado na similaridade entre os usuários, levando em conta 13 característicasextraídas das vendas dos clientes.O resultado mostra que, apesar de chegaram a grupos parecidos, o K-Means teve desempenho levemente superior ao Fuzzy C-Means, no que diz respeito a avaliação através da métrica de estatística Gap.

Palavras-Chave: Clusterização;K-Means; Fuzzy C-Means;

Abstract

This paperanalyzes the application of K-Means and Fuzzy C-Means clustering algorithms.The case study aims to identify customer profiles of an online travel agency, with the objective of improving the effectiveness ofemail marketing campaigns, allowing to send personalized advertisements for each profile.The clustering process was based on similarity among users, considering 13 characteristics extracted from customer sales.The result shows that although they obtained similar groups, the K-Means performed slightly better than Fuzzy C-Means, considering the evaluation through the metric of Gap statistics.

Key-words:*Clustering;K-Means; Fuzzy C-Means;*

http://dx.doi.org/10.25286/repa.v3i3.945

1 Introdução

Uma das técnicas utilizadas durante a mineração de dados é a chamada clusterização (ou agrupamento). O objetivo da clusterização é segmentar determinado conjunto de dados em subgrupos de acordo com similaridades encontradas dentro da base de dados **[1]**. Geralmente essa tarefa é executada de forma não-supervisionada, isto é, sem interferência humana, já que a classificação das amostras é desconhecida.

Um dos algoritmos de clusterização mais tradicionais é o *K-Means*. Ele é bastante difundido devido à sua eficiência e simplicidade, entretanto, o mesmo apresenta alguns problemas quando aplicado em bases de dados mais complexas, sendo os principais: a tendência em convergir para os ótimos locais e o fato de que a escolha dos centroides iniciais interfere bastante na qualidade do agrupamento **[2]**. Tentando corrigir esses problemas, diversos algoritmos alternativos foram sugeridos, dentre eles o *Fuzzy C-Means* (FCM). O FCM faz uso da lógica de agrupamento *fuzzy,* isto é, o conceito de que determinada amostra pode não pertencer a somente um grupo (como no *K-Means*), mas sim a diversos grupos, cada qual com seu grau de pertinência **[3]**.

Este artigo aborda os dois algoritmos. Será feito umestudo de caso utilizando uma base de dados de 2.959 reservas de hotéis de uma agência de viagens online do Brasil durante os anos de 2016 e 2017. A ideia é agrupar os clientes dessas reservas de acordo com similaridades de seus hábitos de compras, utilizando as duas técnicas (*K-Means* e FCM), fazendo um comparativo entre elas, além de realizar uma análise nos agrupamentos de melhor qualidade.

O artigo foi organizado da seguinte maneira: a Seção 2 irá abordar os tipos de clusterização dos algoritmos aqui utilizados,além de apontar alguns trabalhos relacionados. Já na Seção 3 serão detalhados os algoritmos *K-Means*e*Fuzzy C-Means*. A Seção 4 apresenta os detalhes da base de dados utilizada como estudo de caso e a Seção 5 apresenta os resultados. Para finalizar, a Seção 6 apresenta as conclusões sobre estudo de caso e possíveis aplicações futuras.

2 Fundamentação Teórica

2.1 Tipos de Clusterização

A literatura aborda algumas técnicas de clusterização **[4]**, neste artigo vamos apresentar os dois tipos que se aplicam aos algoritmos utilizados no estudo de caso, são eles: particional e *fuzzy*.

O agrupamento particional (do qual o *K-Means* faz parte) tem por objetivo dividir as amostras em grupos (*clusters*, em inglês) que tenham alto grau de similaridade entre seus elementos, e alto grau de separação entre elementos de *clusters* diferentes. Além disso, num algoritmo do tipo particional, cada instância só pode estar atribuída a um *cluster*.

No agrupamento do tipo*fuzzy*, o algoritmo busca dividir as amostras em grupos que podem se sobrepor, isto é, determinada amostra pode pertencer a mais de um cluster. Dada essa natureza, cada amostra apresenta um grau de pertinência em relação a determinado grupo. Caso desejado, um algoritmo *fuzzy* pode ser utilizado para gerar um agrupamento particionalatribuindo determinada amostra ao grupo em que a mesma apresentar o maior grau de pertinência.

2.2 Trabalhos Relacionados

A tarefa de tentar conhecer melhor o comportamento dos passageiros para melhorar a qualidade das indicações de hotéis e pacotes de viagem foi abordada por outros autores. A seguir serão destacados alguns trabalhos que têm essa linha de estudo.

Trabalhos já foram desenvolvidos com o intuito de sugerir pacotes de viagem aos usuários utilizando diversos agentes queinteragem entre si, para buscar as melhores opções de voo, estadia e atrações **[5] [6]**.

http://dx.doi.org/10.25286/repa.v3i3.945

Outros autores avaliaram a utilização de inteligência artificial na área de turismo **[7]**, propondo novos modelos para sistemas de recomendação **[8]** e também elencando ferramentas que já existem com esse propósito.

Também foram encontrados trabalhos que avaliaram técnicas de mineração de dados (filtragem colaborativa e regras de associação) aplicadas ao setor de turismo **[9]**.

Apesar de nenhum dos trabalhos mencionados abordar exclusivamente o tema específico de análise de algoritmos de clusterização numa base de dados de turismo, todos trazem tópicos relevantes no sentido do uso da inteligência artificial para melhorar o relacionamento com os clientes.

3 Algoritmos de clusterização

3.1 *K-Means*

O K-Means é um algoritmo particional proposto porMacQueen em 1967 **[10]**, que é bastante popular por sua simplicidade e eficiência. A ideia do algoritmo é dividir a base de dados em K grupos que tenham instâncias semelhantes, considerando uma medida de similaridade. A similaridade entre duas amostras é medida utilizando uma função de distância, que geralmente é a distância euclidiana. A função de distância euclidiana entre duasamostras x_i e x_j, ambas de dimensão *d (*a dimensão é a quantidade de características de uma amostra), é dada por:

$$d(x_i, x_j) \sqrt{\sum_{p=1}^{d} (x_{i,p} - x_{j,p})^2} \quad \textbf{(1)}$$

Um dos passos mais importantes do *K-Means* é a atualização dos centroides ao fim de cada iteração, através do cálculo das novas médias de cada característica do mesmo. Esse cálculo é feito utilizando a média de cada característica para cada elemento de um determinado. O cálculo é definido através da fórmula:

$$C_k \frac{1}{n_k} . \sum_{i=1}^{n_k} x_i^k \quad \textbf{(2)}$$

Onde C_k representa o centroide do grupo K, en_ko total de amostras presentes no cluster K. Os passos se repetem até que determinado critério de parada seja atingido (por exemplo, quando não houver mais mudança em nenhum dos grupos, ou caso uma quantidade máxima de iterações seja atingida). O Algoritmo 1apresenta o pseudo-algoritmo do *K-Means*.

Algoritmo 1: *K-Means*
Entrada: base de dados com *i* instâncias e *d* dimensões, *K* grupos desejados
Saída: Base de dados dividida em *K* grupos

```
1 início
2 inicializa os K centroides com valores aleatórios;
3 enquanto critério de parada não atingido;
4   para cada amostra x_i faça
5     Adicione x_i ao grupo do centroide C_k de menor distância, de acordo com a equação (1);
6   fim
7   Atualiza os centroides C_k de acordo com a equação (2);
8 fim
```

3.2 *Fuzzy C-Means* (FCM)

O FCM foi introduzido em 1984 por Bezdek **[11]**, como uma extensão do C-Meansparticional. O objetivo do algoritmo é encontrar clusters *fuzzy* para determinado conjunto de dados. A lógica *fuzzy* nesse algoritmo diz respeito ao fato de que, para determinado elemento da base de dados, o algoritmo irá encontrar graus de pertinência para cada cluster, ou seja, um elemento pode pertencer a mais de um cluster.

Para atingir seu objetivo, o FCM segue alguns passos: inicialmente, definimos a quantidade cde clusters desejados, sendo2 ≤ c ≤ n, sendo n a quantidade de amostras.Definimos também o valor do coeficiente de "fuzzificação" m, e iniciamos a matriz de pertencimento $U^{(0)}$ com valores de pertencimento aleatórios. Depois disso, seguimos os seguintes passos:

1. Calculamos os centróides de cada grupo, de acordo com a equação:

http://dx.doi.org/10.25286/repa.v3i3.945

$$v_{ij} = \frac{\sum_{k=1}^{n}(\mu_{ij})^m x_i}{\sum_{k=1}^{n}(\mu_{ij})^m} \quad \textbf{(3)}$$

2. Calcula a distância euclidiana D_{ij}, utilizando a equação (1) de cada ponto ipara cada centroidej.
3. Atualiza os valores μ_{ij} da matriz de pertencimento U, seguindo a equação:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{Dkj})^{\frac{2}{m-1}}} \quad \textbf{(4)}$$

O algoritmo executa esses passos até que o módulo da diferença entre duas matrizes de pertencimento U^k e U^{k-1} seja menos que o coeficiente de erro ε definido pelo usuário. Essa condição é definida na equação:

$$\|U^k - U^{k-1}\| < \varepsilon \quad \textbf{(5)}$$

Outros critérios de paradas também podem ser adotados, como por exemplo: atingir um número máximo de interações. Sendo assim, poderíamos definir o pseudo-algoritmo do *Fuzzy C-Means* da seguinte maneira:

Algoritmo 2: *Fuzzy C-Means (FCM)*

Entrada: coeficiente de "fuzzificação" m, *c* grupos desejados, base de dados com i instâncias, ε erro aceitável.

Saída: Matriz de pertencimento M com c colunas e i linhas, onde o elemento M_{ic} define o grau de pertencimento do elemento i no conjunto c.

1. **início**
2. inicializa a matriz de pertencimento M com valores aleatórios entre 0 e 1;
3. **repita**;
4. cacula os centroides de acordo com a equação (3), para cada cluster c.
5. cacula a distância euclidiana de cada instância, para cada centroide de acordo com a equação (1)
6. Atualiza a matriz de pertinência de acordo com a equação (4);
7. **até que** equação (5) seja verdadeira;
8. **fim**

4 Base de Dados de Reservas de Hotéis

Uma agência de viagens online, possui um *site* de reservas de hotéis, onde oferece hospedagens para hotéis em todo o Brasil.Os clientes, em sua grande maioria, são atingidos através de disparos de e-mail marketing com ofertas determinadas de maneira arbitrária pelo departamento de marketing da agência, sem nenhum tipo de filtro sobre os destinatários da oferta. Essa abordagem acarreta um grande custo com os envios do e-mail marketing, uma vez que o custo é determinado pela quantidade de e-mails destinatários.

Visando melhorar a eficiência do uso do dinheiro investido em e-mail marketing, foi pensando em realizar uma classificação dos usuários do *site*, com base em seu histórico de compras, com o objetivo de identificar perfis de clientes e, dessa maneira, enviar oferta de maneira mais eficaz, ou seja, apenas aos usuários com mais probabilidade de se interessar pelo anuncio.

Para realizar o agrupamento dos usuários, foram selecionadas 2.959 vendas realizadas pelo site da agência. Essas vendas foram escolhidas devido ao fato de ser o total de vendas aprovadas durante os anos de 2016 e 2017, já os foram

http://dx.doi.org/10.25286/repa.v3i3.945

escolhidos pois a empresa focou nas vendas online de viagens a partir de 2016.

As amostras são compostaspor 13 características, descritas na Tabela 1. Todos os valores das amostras foram normalizados para uso nos algoritmos.

Tabela 1 - Características de cada elemento da amostra

Característica	Descrição
Uf	Estado onde o cliente reside
Idade	Idade do clientes
Total Comprado	Soma total de todas as compras do cliente
Ticket Médio	Valor médio de compra do cliente
Quantidade de Compras	Quantidade de compras do cliente
Tipo de Pagamento	Forma que o cliente pagou a compra (Cartão de crédito, boleto)
Quantidade de Parcelas	Quantidade de parcelas que o cliente escolheu dividir
Cidade	Cidade do cliente
Valor da venda	Valor da venda específica
Hotel	Hotel comprado pelo cliente
Destino	Cidade de destino da viagem
Mês de estadia	Mês que o cliente escolheu se hospedar
Quantidade de diárias	Total de dias que o cliente irá ficar hospedado

Tabela 2 - Resultado da simulação dos algoritmos utilizando 30 simulações e condição de parada 1.000 iterações.

Algoritmo-K	Estatística GAP	Distância Inter-cluster	Erro Quantizado	Distância Intra-Cluster
K-Means-2	0.14(0.05)	0.58(0.03)	0.53(0.008)	1595.98(8.13)
K-Means-3	**0.25(0.07)**	2.14(0.08)	0.50(0.014)	1480.49(28.91)
K-Means-4	0.31(0.07)	4.77(0.32)	0.47(0.010)	1407.52(10.05)
K-Means-5	0.37(0.06)	8.49(0.51)	0.46(0.013)	1354.52(10.16)
K-Means-6	0.44(0.07)	13.81(0.66)	0.45(0.011)	1306.17(15.31)
K-Means-7	0.48(0.07)	19.76(0.98)	0.44(0.006)	1264.06(14.54)
K-Means-8	0.49(0.08)	27.84(0.90)	0.43(0.006)	1233.99(18.03)
K-Means-9	0.54(0.09)	36.94(1.67)	0.42(0.005)	1205.38(12.19)
K-Means-10	0.56(0.07)	**47.34(2.19)**	**0.41(0.004)**	**1183.41(6.33)**
FCM-2	0.12(0.04)	0.05(0.014)	0.569(0.005)	1690.90(21.83)
FCM-3	**0.21(0.05)**	0.24(0.002)	**0.564(0.006)**	1645.34(2.00)
FCM-4	0.19(0.04)	0.55(0.003)	0.575(0.017)	1624.03(1.51)
FCM-5	0.20(0.06)	0.91(0.003)	0.581(0.030)	1615.03(1.33)
FCM-6	0.17(0.06)	1.30(0.004)	0.563(0.018)	1611.44(1.75)
FCM-7	0.16(0.06)	1.72(0.004)	0.564(0.021)	**1609.54(1.65)**
FCM-8	0.17(0.07)	2.17(0.004)	0.597(0.020)	1609.94(2.10)
FCM-9	0.13(0.07)	2.65(0.004)	0.585(0.036)	1610.99(2.36)
FCM-10	0.12(0.07)	**3.17(0.005)**	0.593(0.030)	1612.85(3.36)

http://dx.doi.org/10.25286/repa.v3i3.945

4.1 Pré-processamento dos Dados

Com o intuito de melhorar a qualidade dos dados trabalhados, alguns pré-processamentos foram feitos em algumas características. Todos os tratamentos foram feitos utilizando a ferramenta de códigoaberto *Open* Refine **[12]**. As características que sofreram algum tipo de manipulação foram:

1) Idade: característica extraída através da data de nascimento do cliente.
2) Cidade: como no site da agência essa característica é um campo aberto para digitação do usuário, foi necessário padronizar o nome das cidades para evitar que registros da mesma cidade aparecessem com valores distintos.
   Exemplo: São Paulo, São Paulo e Sampa foram unificados para São Paulo, e assim por diante.
3) Destino: alguns hotéis de um mesmo destino constavam como destinos diferentes. Exemplo:

   dois hotéis A e B ficam em Porto de Galinhas/PE, mas o hotel A tinha como destino Ipojuca/PE. Como Porto de Galinhas é uma praia de Ipojuca, foram unificados para Porto de Galinhas (sempre unificamos para o nome que o destino é mais conhecido).
   Mês de estadia: característica foi derivada com base no dia da entrada do cliente no hotel, isto é, se a estadia de um cliente começou em 01/04/2018, o mês de estadia foi abril, mesmo que a estadia tenha se estendido por meses distintos.

5 Experimento e Resultados

O estudo de caso foi realizado utilizando a seguinte estratégia: como ambos os algoritmos têm como entrada uma quantidade K de clusters desejados, foram realizadas 30 execuções do algoritmo, para cada quantidade K de clusters, com o K variando de 2 a 10 grupos. Posteriormente, foi realizada uma média de 4 métricas para cada valor de K relativas às 20 execuções, de maneira que avaliando essas métricas, fosse possível definir qual o melhor K a ser escolhido para cada algoritmo. O resultado consolidado das execuções está presente na Tabela 2.

As métricas utilizadas para avaliar os grupos foram as seguintes: Estatística GAP, Distâncias Intra e Inter-cluster e o Erro quantizado. A seguir explicamos cada uma dessas métricas e, posteriormente, iremos avaliar os grupos que tiveram a melhor estatística GAP para ambos os algoritmos, pois essa métrica se mostra eficiente na decisão de escolher a melhor quantidade de grupos **[13]**.

5.1 Descrição das métricas

Estatística GAP: é uma métrica proposta por Tibshirani **[13]**, em 2001, que tem por objetivo encontrar o número ideal de clusters K. Para escolha do K ideal, são avaliados os diversos valores da métrica (que faz uso do logaritmo da distância intra-cluster) para os diversos valores de K. Essa mesma análise também é feita para um conjunto de dados aleatórios. A estatística GAP representa justamente a diferença entreo valor encontrado para a amostra real em relação à amostra aleatória. Por isso o valor dessa métrica deve ser maximizado, mostrando que o agrupamento escolhido se difere de um agrupamento aleatório. Para o cálculo dessa métrica, utilizam-se as seguintes equações:

$$D_k = \sum_{\forall xi,xj \in Ck} d(x_i, x_j) \quad \textbf{(6)}$$

Onde d(x_i,x_j) é a distância euclidiana entre os pontos. Posteriormente, utilizamos a equação (6) para encontrar a dispersão entre valores crescentes de K, através da equação:

http://dx.doi.org/10.25286/repa.v3i3.945

$$W_k = \sum_{i=1}^{k} \frac{1}{2n_r} D_i \quad \textbf{(7)}$$

Uma vez que temos a dispersão interna dos grupos, utiliza-se a versão amortizada logW_k com o valor da mesma métrica para uma amostra de dados aleatória. Sendo assim, a estatística GAP tem por objetivo maximizar o valor da equação:

$$Gap_n(\mathrm{k}) = E_n^*(\log W_k) - \log W_k \quad \textbf{(8)}$$

Onde E_n^* (log W_k)representa o valor da métrica esperado para uma amostra aleatória. Distância Intra-Cluster: é utilizada para validar a distância entre dois elementosx_i e x_j,pertencentes ao mesmo grupoC_k. Sendo assim, um bom valor para essa métrica é um valor baixo, indicando proximidade entre os elementos do cluster. Ela é medida conforme a equação:

$$D_{intra} = \sum_k \frac{1}{2N_k} \sum_{xi,xj \in Ck} d(x_i, x_j) \quad \textbf{(9)}$$

Distância Inter-Cluster: é utilizada para validar a distância entre dois centroides C_k e $C_{K'}$. Sendo assim, um bom valor para essa métrica é um valor alto, indicando que os clusters estão bem separados. É calculada através da equação:

$$D_{inter} = \sum_{\forall k,k' | k \neq k'} d(c_k, c_{k'}) \quad \textbf{(10)}$$

Erro Quantizado:é utilizada para medir a eficiência do algoritmo para valores crescentes de K. Leva em consideração a distância euclidiana de cada ponto x_i ao seu centroide c_k em relação ao total de amostras pertencentes ao cluster C_k , referenciado na fórmula como $|C_k|$. Sendo assim, o valor ideal para essa métrica são valores baixos. Segue a equação:

$$J_e = \frac{\sum_k \sum_{\forall x_i \in C_k} d(x_i, c_k)/|C_k|}{N_k} \quad \textbf{(11)}$$

5.2 Resultados *K-Means*

Os resultados da execução do *K-Means*, estão exibidos na Tabela 2. Através dela, podemos perceber que o algoritmo conseguiu cumprir o que se esperava dele, ao maximizar a distância inter-cluster e minimizar a distância intra-cluster, conforme o numero de K aumenta. Também podemos perceber que o algoritmo também conseguiu diminuir a métrica do erro quantizado, que também leva em consideração a qualidade interna dos grupos. Analisando os valores da estatística Gap, também percebemos sucesso ao maximizar o valor da métrica. Nas figuras 1, 2, 3 e 4 podemos ver a variação das métricas das distâncias intra e inter cluster, o erro quantizado e a estatística gap, respectivamente.

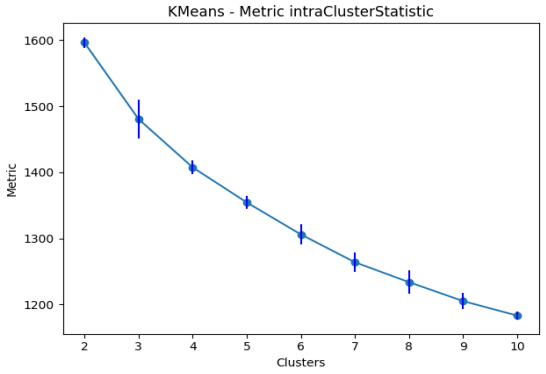

Figura 1 - Variação da distância intra-cluster para valores crescentes de K, utilizando K-Means.

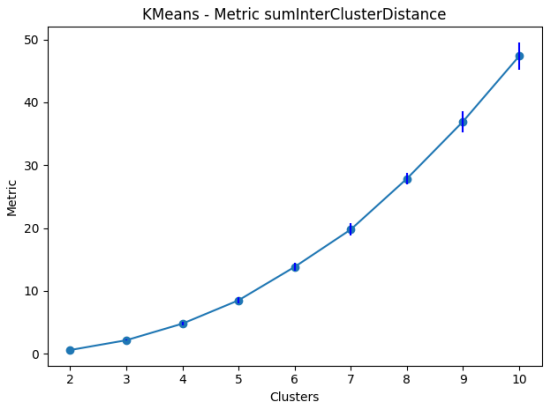

Figura 2 - Variação da distância inter-cluster para valores crescentes de K, utilizando *K-Means*.

http://dx.doi.org/10.25286/repa.v3i3.945

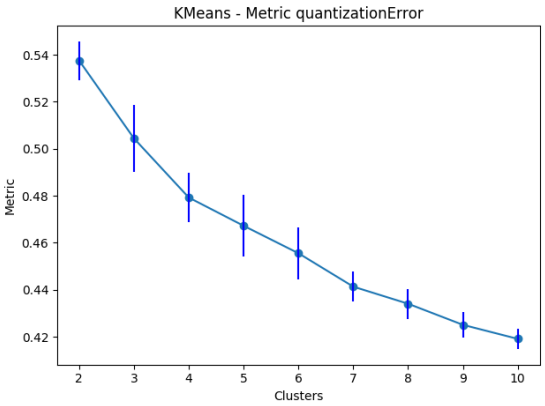

Figura 3 - Variação do erro quantizado para valores crescentes de K, utilizando *K-Means*.

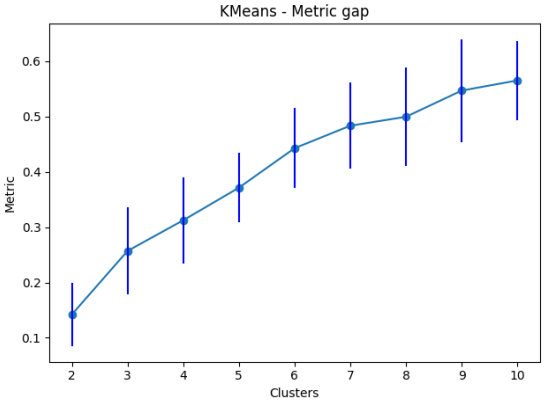

Figura 4 - Variação da estatística Gap para valores crescentes de K, utilizando *K-Means*.

Com relação ao número de clusters ideal de acordo com cada métrica, podemos analisar que as métricas das distâncias intra e inter-cluster e a do erro quantizado, apontaram o número de grupos 10 como ideal. Já a estatística gap, apontou como quantidade ideal 3. Apesar de três métricas apontarem 10 como quantidade ideal, podemos indicar 3 como quantidade de grupos a ser escolhida, uma vez que as distâncias intra e inter-cluster têm a tendência de se otimizarem com a quantidade de K aumentando. A quantidade 3 também foi apontada como ideal pela equipe de negócio do site de viagens, uma vez que uma granularização de 10 clusters não seria tão necessária para o direcionamento das ações de e-mail marketing.

Posteriormente, com o intuito de identificar o perfil de usuários criados pelo *K-Means*, com 3 grupos, foi analisada uma das amostras agrupadas a fim de identificar quais características pesaram na escolha dos clusters:foi percebido que o *K-Means* criou os clusters principalmente considerando a variável "Mês de Estadia", sendo um cluster de usuários que se hospedam mais nos primeiros meses do ano, outro grupo de usuários que se hospedam mais próximo ao meio do ano e o último mais próximo do final do ano.

5.3 Resultados *Fuzzy C-Means* (FCM)

Os resultados da execução do *Fuzzy C-Means*, estão exibidos na tabela 2. Podemos perceber que ao contrário do que ocorreu no *K-Means*, as métricas de erro quantizado e a distância intra-cluster, não mantiveram a minimização conforme a quantidade de grupos aumentou.

Podemos perceber que para a distância intra-cluster, a minimização funcionou bem até o k igual 7, depois disso o índice começou a subir. A dificuldade de encontrar grupos coesos pode ser percebida também na oscilação da métrica do erro quantizado, que ficou oscilando para cada número de k.Em relação à distância intra-cluster, o algoritmo demonstrou o comportamento esperado, maximizando a distância entre os grupos, conforme aumenta a quantidade de clusters. Já a estatística Gap, não manteve um aumento durante os testes com número crescente de K, ela atingiu o valor ótimo com k igual a 3, e depois teve uma pequena oscilação até diminuir no k igual a 10, indicando perda de qualidade nos agrupamentos.Nas figuras 5, 6, 7 e 8, podemos ver a variação das métricas das distâncias intra e inter cluster, o erro quantizado e a estatística gap, respectivamente.

http://dx.doi.org/10.25286/repa.v3i3.945

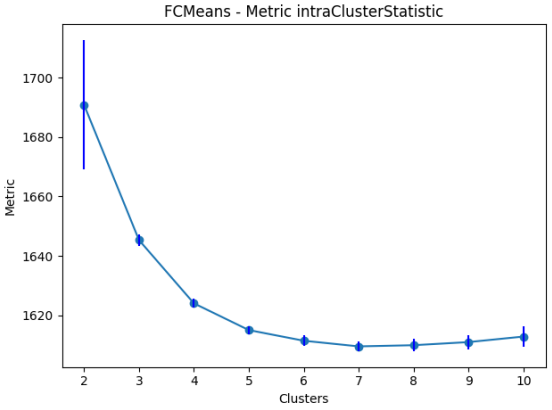

Figura 5 - Variação da distância intra-cluster para valores crescentes de K, utilizando FCM.

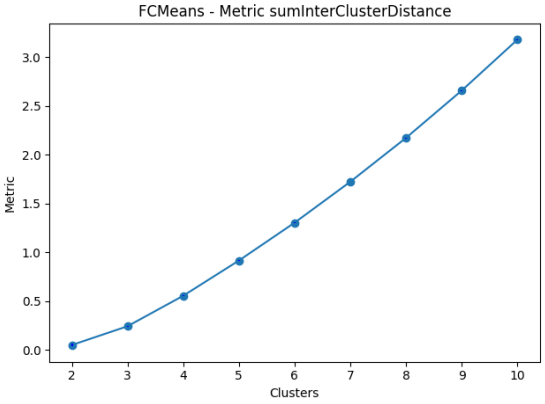

Figura 6 - Variação da distância inter-cluster para valores crescentes de K, utilizando FCM

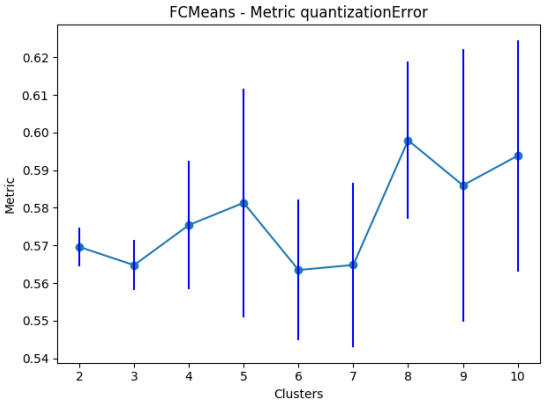

Figura 7- Variação do erro quantizado para valores crescentes de K, utilizando FCM.

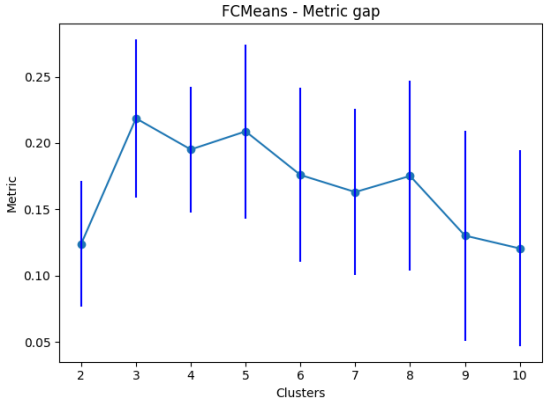

Figura 8- Variação da estatística Gap para valores crescentes de K, utilizando FCM.

Com relação à quantidade de grupos ideal, podemos perceber que as métricas do erro quantizado e estatística gap, apontaram a quantidade de grupos igual a 3 como melhor escolha (apesar do erro quantizado do k igual 7 possui o valor da métrica igual ao 3, o desvio padrão é maior, fazendo com que predominasse o k igual a 3 para essa métrica). As distâncias intra e inter-cluster, apontaram o k igual a 7 e 10, respectivamente.

Para identificação do perfil dos usuários criados pelo FCM com 3 grupos, também foi analisada uma amostra agrupada com o intuito de identificar características que se destacam entre os perfis: também foi identificado que a característica "Mês da estadia" puxou a criação dos grupos, de maneira igual ao K-Means, com usuários que se hospedam no início, meio e fim do ano.

5.4 Análise e Discussão

Durante a análise e comparação dos resultados dos algoritmos, pudemos perceber que existiram duas características que também poderiam definir os perfis dos usuários, foram elas o Destino e o Hotel. Quando essas variáveis foram transformadas de categóricas para numéricas, o processo utilizado foi o de ordená-las em ordem alfabética e atribuir números sequênciais, iniciando em 1. Exemplificando: supondo que tivéssemos 3 cidades: Atibaia, Bauru e Campo Grande.Seguindo a lógica utilizada, elas seriam transformadas em 1, 2 e 3, respectivamente.

http://dx.doi.org/10.25286/repa.v3i3.945

Devido à forma como essas características foram transformadas, o ganho de informação não foi relevante, isto porque os grupos em que os usuários se dividiram seguiram a ordem alfabética das características, o que apontou, por exemplo, grupos em que determinado usuário compra mais hotéis/destinos com as letras do começo do alfabeto.

Como uma forma de melhorar esse cenário, poderíamos trabalhar essas duas características para gerar uma terceira, que seria o tipo de hotel. A lógica seria, tendo como base o Hotel e o Destino, apontar qual o perfil em que o hotel se encaixaria melhor (Praia, Campo, entre outros). Com essa terceira característica, o modelo poderia conseguir uma melhor generalização, apontando perfis de usuário por tipo de hotel.

6 Conclusões

Este artigo fez uma análise da aplicação de dois algoritmos de agrupamento (*K-Means* e *Fuzzy C-Means*) em uma base de dados de reservas de hotéis de uma agência de viagens online do Brasil. As características do conjunto de dados eram detalhes sobre as vendas efetuadas no site da agência durante os anos de 2016 e 2017. O objetivo foi identificar perfis de clientes com o intuito de melhorara eficácia do envio de ofertas através de e-mail marketing, possibilitando o envio de anúncios personalizados para cada perfil.

Foram escolhidos o *K-Means* e o *Fuzzy C-Means* por serem alguns dos algoritmos mais populares na aplicação de agrupamento em bases de dados, além do fato, de cada um utilizar abordagens de clusterização diferentes (*K-Means* é particionista e o FCM utiliza lógica *fuzzy*).

Os resultados mostraram que ambos os algoritmos chegaram a um número de K ideal igual a 3, considerando a métrica de avaliação de clusters Estatística Gap. Considerando os dados agrupados, os dois algoritmos também tiveram resultados semelhantes, dando ênfase à característica "Mês de Estadia" para criar os grupos. Apesar da similaridade nos resultados, podemos apontar que o *K-Means* teve desempenho levemente superior, uma vez que o valor da sua estatística Gap foi maior que a do FCM, demonstrando grupos mais consistentes.

Como trabalhos futuros, podemos agregar outros algoritmos de clusterização para serem analisados. Também podemos aplicar a clusterização resultada desse estudo de caso, para ser validada na prática, avaliando métricas de envio de campanhas por e-mail como a taxa de abertura das campanhas segmentadas em relação com as das campanhas sem segmentação.

Referências

[1] PROVOST, Foster; FAMCETT, Tom. Data science and its relationship to big data and data-driven decision making. **Big Data**, v. 1, n.1 , p. 51-59, 2013. Disponível em: <https://www.liebertpub.com/doi/abs/10.1089/big.2013.1508>.

[2] MUMTAZ, Karam; DURAISWAMY Karthig. A Novel Density based improved k-means Clustering Algorithm – Dbkmeans. **International Journal on Computer Science and Engineering**, v. 2, n.2, p. 213-218, 2010. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.5204&rep=rep1&type=pdf>

[3] GHOSH, Soume; DUBEY, Sanajy Kumar. Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. **International Journal of Advanced Computer Science and Applications**, v. 4, n.4, p. 35-39, 2013. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.683.5131&rep=rep1&type=pdf#page=46>.

[4] JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data Clustering: A Review. **ACM Computing Surveys**, v. 31, n.3, p. 264-323, 1999. Disponível em: <https://dl.acm.org/citation.cfm?id=331504>.

http://dx.doi.org/10.25286/repa.v3i3.945

[5] LORENZIi, Fabiana et al. Enhancing the Quality of Recommendations Through Expert and Trusted Agents. In: IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, 23., 2011, Florida. **Proceedings**... Florida: IEEE, 2011. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6103346/>

[6] Lorenzi, Fabiana; LOH, Stanley; ABEL, Mara. PersonalTour: a recommender system for travel packages. In: IEEE/WIC/ACM INTERNACIONAL CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECNOLOGY, 4., 2011, Lyon. **Proceedings**... Lyon: IEEE, 2011. p. 333-336. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6040800/>.

[7] COELHO, Bruno; MARTINS, Constantino; ALMEIDA, Ana. Web Intelligence in Tourism: user modeling and recommender system. In IIEEE/WIC/ACM INTERNACIONAL CONFERENCE ON WEB INTELLIGENCE AND INTELLIGENT AGENT TECNOLOGY, 3., 2010, Califórnia. **Proceedings**... Califórnia: IEEE, 2010. p. 619-622. Disponível em: <https://ieeexplore.ieee.org/abstract/document/5616446/>.

[8] SANTOS, Filipe et al. Tourism Recommendation System based in user's profile and functionality levels. In: INTERNATIONAL C* CONFERCENCE ON COMPUTER SCIENCE & SOFTWARE ENGINEERING, 9., 2016, Porto. **Proceedings**... Porto: ACM, 2016. p. 93-97. Disponível em: <https://dl.acm.org/citation.cfm?id=2948995>.

[9] ZHAO, Xuesong; JI, Kkaifan. Tourism E-Commerce Recomender System Based on Web Data Mining. In: The International Conference on Computer Science & Education, 8., 2013, Sri Lanka. **Proceedings**... Sri Lanka: IEEE, 2013. Disponível em: <https://ieeexplore.ieee.org/abstract/document/6554161/>.

[10] MacQueen, James. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, Califórnia, v. 1, n.14, p. 281-297, 1967. p. 281-297.

[11] BEZDEK, James C.; EHRLICH, Robert; FULL, Willam. FCM: The Fuzzy C-Means Clustering Algorithm. **Computers & Geosciences**, v. 10, n.2-3, p. 191-203, 1984.

[12] Open Refine. Disponível em: <http://openrefine.org>. Acessoem: 23 abr. 2018.

[13] TIBSHIRANI, Roberto; GUENTHER, Walther; HASTIE, Trevor. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society**: Series B, v. 63, Parte 2, p. 411-423, 2001.

http://dx.doi.org/10.25286/repa.v3i3.945