Estudo Comparativo dos Algoritmos Fuzzy C-Means e K-Means aplicados à segmentação de consumidores

Amanda Lucas Pereira

Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brasil
amandalucaspereira@gmail.com

Resumo O presente trabalho aplica o algoritmo Fuzzy C-Means em um conjunto de dados de consumo de clientes de uma dada empresa, buscando encontrar uma forma satisfatória de particionar os mesmos. Adicionalmente, é realizada uma análise da influência do valor do coeficiente nebuloso no ajuste do modelo aos dados. Código está disponibilizado no github.

1 Introdução

A análise do perfil de clientes de uma empresa é crucial para a manutenção e conquista de nova clientela. Nesse contexto, algoritmos de agrupamento podem ser empregados visando a extração e interpretação de informações relevantes referentes às partições que formam o conjunto de clientes de uma empresa. Por se tratarem de dados que apresentam certa sobreposição entre grupos, a aplicação de métodos que realizam *soft clustering* se torna interessante e um dos algoritmos popularmente testados em tais aplicações é o Fuzzy C-Means (FCM). O presente trabalho busca realizar um estudo comparativo do FCM com o algoritmo K-Means tentando atacar o problema de segmentação de consumidores utilizando um conjunto de dados referente ao consumo de clientes cadastrados em um empresa.

O restante do trabalho está distribuído em mais sete seções: na seção 2 é apresentada uma breve introdução dos fundamentos técnicos necessários para melhor compreensão dos métodos e modelos desenvolvidos neste trabalho. A metodologia utilizada para solução do problema proposto, assim como sua aplicabilidade à esfera do problema é descrita na terceira seção. Os experimentos que foram realizados com os algoritmos e os seus respectivos resultados são discutidos na seção 4 e 5. Na seção 6, é realizada uma análise comparativa dos resultados obtidos. Por fim, a última seção encerra o trabalho apresentando as conclusões e perspectivas de novos trabalhos.

2 Referencial Teórico

2.1 Algoritmos de Agrupamento

O grupo formado pelos algoritmos de agrupamento (*clustering*, ou clusterização) compreende diferentes métodos de aprendizado que são aplicados em um conjunto de dados com o intuito de agrupar objetos que sejam similares em uma mesma partição e analogamente, objetos dissimilares em partições distintas [7]. A medida de similaridade utilizada varia de acordo com a aplicação e com o método escolhido, podendo ser definida como a distância entre objetos ou funções de similaridade [6].

Os métodos conhecidos como sendo do tipo *hard clustering* dividem os objetos de um conjunto de dados em grupos disjuntos. Ou seja, cada objeto pertence unicamente a um único *cluster*. Os métodos que realizam *soft clustering* estendem essa ideia de forma que um mesmo objeto possa pertencer a diferentes *clusters*.

2.2 K-Means

Um dos algoritmos de agrupamento mais amplamente usados é o K-Means. Nesse algoritmo, o conjunto de dados é particionado em k grupos, sendo que cada grupo tem como seu centro a média das amostras pertencentes a este. O parâmetro k que define o número de clusters deve ser definido pelo especialista [3].

O processo de otimização ocorre da seguinte forma: inicializa-se os centros dos clusters seguindo algum tipo de procedimento; a cada iteração, as amostras são atribuídas ao cluster cujo centro está mais próximo da mesma; os centros dos clusters são recalculados. O ciclo é repetido até que algum critério de convergência seja cumprido, que geralmente é uma medida de erro [6].

2.3 Fuzzy C-Means

O algoritmo Fuzzy C-Means (FCM) é um método do tipo *soft clustering*, o qual atribui para cada objeto valores diferentes de pertinência. Os valores de pertinência de uma amostra são referentes a cada *cluster*, de forma que um mesmo objeto possa pertencer a um ou mais grupos [6,1]. A atribuição dos valores de pertinência é realizada por meio de uma função de membresia, que mede a pertinência de uma dada amostra a um conjunto fuzzy.

Seja c o número de clusters que se deseja obter, e n o número de amostras no conjunto de dados. O FCM busca minimizar uma função objetivo dada por

$$L = \sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^{m} ||x_j - v_i||^2 \tag{1}$$

onde u_{ij} representa a medida de membresia da amostra x_j ao cluster i e v_i se refere ao centro do i-ésimo cluster. Representado por m, o coeficiente nebuloso é um parâmetro que controla a nebulosidade da partição resultante [2].

Estimativa do coeficiente nebuloso O coeficiente nebuloso é um parâmetro que determina, como o próprio nome indica, a nebulosidade dos clusters obtidos pelo FCM. Um coeficiente no valor de 1 é equivalente a *hard clustering*, e valores maiores que 1.0 costuma abaixar os valores de pertinência das amostras nas bordas dos clusters.

Este parâmetro pode ser ajustado de diferentes maneiras, como a aplicação de um algoritmo genético. Para o presente trabalho, foi utilizado uma equação apresentada pelos autores em [8]:

$$\begin{aligned} f(D,n) = 1 + \left(\frac{1418}{n} + 22.05\right) D^{-2} \\ + \left(\frac{12.33}{n} + 0.243\right) D^{0.0406ln(n)0.1134} \end{aligned} \tag{2}$$

Onde n se refere ao número de amostras e D a dimensão do conjunto de dados.

Coeficiente de Partição Alguns tipos de funções que medem a validade dos clusters obtidos ao final do FCM podem ser utilizadas. O Coeficiente de Partição (PC) é uma métrica indicativa de quão bem um conjunto de dados está sendo descrito por um modelo treinado, geralmente calculada a cada iteração do FCM [1]. O *range* dessa métrica é de 0 a 1, sendo 1 o melhor. Modelos FCM treinados com diferentes parâmetros – número de clusters, coeficiente nebuloso, entre outros – podem ter sua performance comparada avaliando essa métrica em conjunto com o Coeficiente de Entropia.

Coeficiente de Entropia O coeficiente de entropia (PE) é uma medida da nebulosidade apresentada nas partições propostas pelo FCM. O PE tem seu range de $[0, log(C)]$ e quanto mais próximo de 0, o mais crisp os conjuntos de saída são [4].

2.4 Principal Component Analysis

O método chamado *Principal Component Analysis* é aplicado em um conjunto de dados com o objetivo de identificar novas variáveis, chamadas de componentes principais, formadas por combinações lineares dos atributos originais. Após aplicado o PCA, a primeira componente principal se refere à direção para qual os dados apresentam maior variância. Para a segunda componente, a segunda maior variância, e assim em diante.

Esse método possibilita a redução de conjunto de dados, possibilitando a execução de algoritmos usando uma quantidade inferior de atributos mantendo um resultário coerente com a utilização dos dados originais.

3 Metodologia

3.1 Conjunto de dados

O conjunto de dados utilizado para o presente trabalho é o "Customer Personality Analysis"e está disponbilizado na plataforma Kaggle [5]. O conjunto consiste em 2240 amostras, e apresenta informações relativas aos clientes de uma dada empresa.

As informações presentes no conjunto de dados referentes aos dados pessoais de um cliente são: um ID de identificação do cliente; o ano de nascimento; nível educacional; um atributo relativo ao estado civil; salário; número de filhos (crianças); número de filhos adolescentes; data de cadastro do cliente; contagem de dias desde a última compra; e uma variável binária que indica se o cliente realizou alguma reclamação nos últimos dois anos.

Os atributos que tratam de produtos consumidos pelos clientes são referentes aos últimos dois anos, e são os seguintes: quantia gasta na compra de vinhos; quantia gasta em frutas; quantia gasta em carne; quantia gasta em peixe; quantia gasta em doces; quantia gasta em ouro.

Adicionalmente, o conjunto traz variáveis relacionadas ao consumo de ofertas pelos clientes, sejam essas por meio de desconto ou de campanhas especiais. São essas variáveis: número de compras com desconto; variáveis binárias definindo se o cliente aceitou a oferta em uma campanha de número X, com X entre 1 e 5; e uma variável que indica se o cliente aceitou a oferta oferecida em uma última campanha.

Os últimos atributos presentes no conjunto utilizado são referentes ao local de compra dos consumidores: número de compras realizadas pelo website da empresa; número de compras realizadas via revista; número de compras em lojas físicas; e número de visitas ao website da empresa no último mês.

3.2 Pré-processamento dos dados

A etapa inicial do pré-processamento consiste na limpeza do conjunto de dados. Primeiramente, foram descartadas as amostras que apresentassem algum valor inadequado para algum atributo – como a presença de "NaN". Em seguida, realizando uma análise do atributo referente à idade dos consumidores, notou-se que existia registros referentes a pessoas com mais de 100 anos de idade; e de forma análoga, percebeu-se a existência de *outliers* na variável referente à salário (Figura 1). Essas amostras também foram descartadas.

Em seguida, foram aplicadas transformações em alguns atributos para que todos se encontrassem da forma ideal para o processamento pelo algoritmo de clusterização: dados numéricos e normalizados. Adicionalmente, usando a informação da data de cadastro dos consumidores, construiu-se um novo atributo: tempo total de cadastro do consumidor nos registros da empresa. Também foram construídas as seguintes variáveis: total de filhos (dado pela soma dos filhos crianças e adolescentes); um atributo indicando se a pessoa mora sozinha ou não; e um atributo referente ao gasto total em produtos.

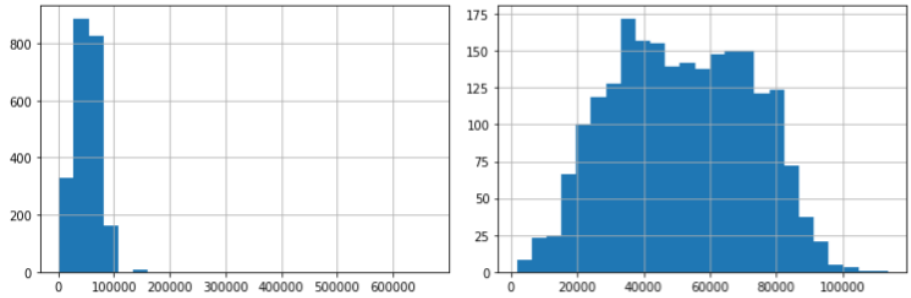

Figura 1. Análise visual para remoção de *outliers* referente ao salário dos consumidores. À esquerda: antes da filtragem, à direita: após filtragem.

Após a limpeza e criação de atributos, o conjunto de dados compreendia 2205 amostras com 32 atributos referentes aos clientes. Os dados foram divididos entre dados de treino e teste: 1764 para ajustes do modelo e 441 para avaliação do mesmo. A última etapa do pré-proceesamento foi a normalização de cada atributo. Para isso, utilizou-se a média e o desvio padrão de cada atributo nos dados de treinamento. Em seguida, os dados de teste foram normalizados utilizando as mesmas estatísticas descritivas encontradas nos dados de treinamento.

4 Experimentos com o Fuzzy C-Means

As seções seguintes apresentam as análises realizadas para diferentes experimentos. Inicialmente, o FCM é aplicado no conjunto de atributos chamado de *baseline*, que compreende os dados com os 32 atributos normalizados citados nas seção anterior. Em seguida, são apresentados três experimentos utilizando PCA. Para cada experimento, foram realizadas as mesmas etapas relativas à aplicação do FCM. Adicionalmente, é realizada uma avaliação da influência de diferentes valores do coeficiente nebuloso m em um mesmo contexto de análise.

Para cada configuração apresentada, foram testados diferentes números de clusters: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14. A definição do número ideal de clusters foi realizada por análise gráfica dos parâmetros de Coeficiente de Partição (PC) e Coeficiente de Entropia (PE). Primeiramente, foram realizados experimentos usando apenas o FCM para definição do número de clusters ideal para o conjunto de dados. Em seguida, foram realizados experimentos com o K-Means para fim de comparação entre os algoritmos.

4.1 Conjunto de atributos original

Usando a equação apresentada 2, o parâmetro m para o conjunto de atributos original foi estimado como $m_{est} = 1.088$. Analisando os gráficos de PE e PC, um bom número de clusters para essa configuração parece ser $c = 4$ (Figura 11).

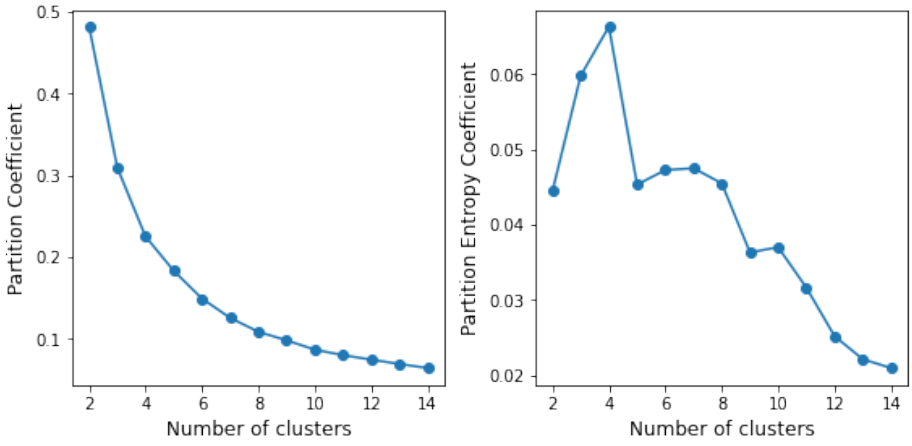

Figura 2. Coeficientes obtidos para cada número de clusters utilizando o conjunto de atributos original.

4.2 Conjunto de atributos aplicando *Principal Component Analysis*

Foram realizados testes para avaliar a influência da utilização de PCA para o FCM, variando o número de componentes principais entre 2, 3 e 5. Os resultados estão apresentados nas Figura 3, em conjunto com a curva obtida para o conjunto de dados apresentado anteriormente. Os valores estimados do coeficiente m para cada configuração foram: $m_{est,2} = 6.90$, $m_{est,3} = 3.697$ e $m_{est,5} = 2.041$.

Os melhores valores de PC e PEC foram obtidos com os dados originais. Para gerar uma comparação com o baseline, foi escolhido o método que utiliza 3 componentes principais, por questão de facilitar a visualização e não apresentar resultado tão inferior quanto o que mais se aproxima do baseline – $p = 5$. Analisando as curvas, um bom número de clusters para esse conjunto de dados parece ser $c = 2$ ou $c = 4$. Visto que dois clusters parece ser um valor relativamente baixo, optou-se por seguir os experimentos utilizando $c = 4$.

4.3 Testes da influência do coeficiente nebuloso

Utilizando o conjunto de atributos formado por 3 componentes principais, foi realizada uma análise da influência de variar o valor do coeficiente nebuloso m (Figura 4). A escolha do coeficiente parece apresentar maior influência nos valores de PC e PE obtidos quando se utilizou um maior número de clusters, com $c = 4$, em comparação ao experimento com $c = 2$.

5 Experimentos com o K-Means

Os experimentos com o K-Means foram realizados com o parâmetro $k = 4$ para que seja possível realizar uma análise comparativa com o FCM. Por inspeção

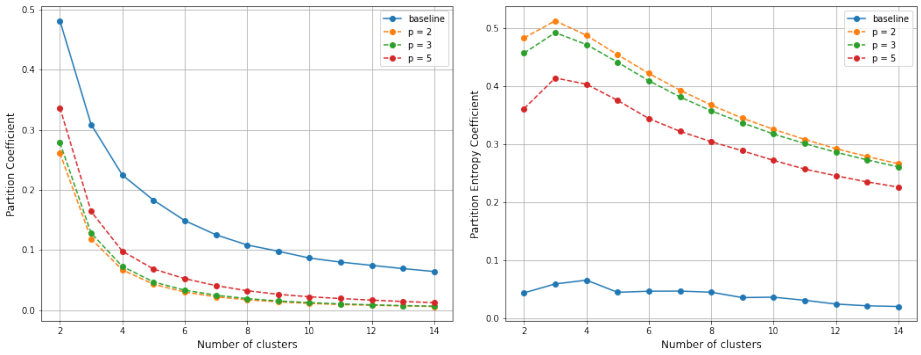

Figura 3. Coeficientes obtidos para cada número de clusters.

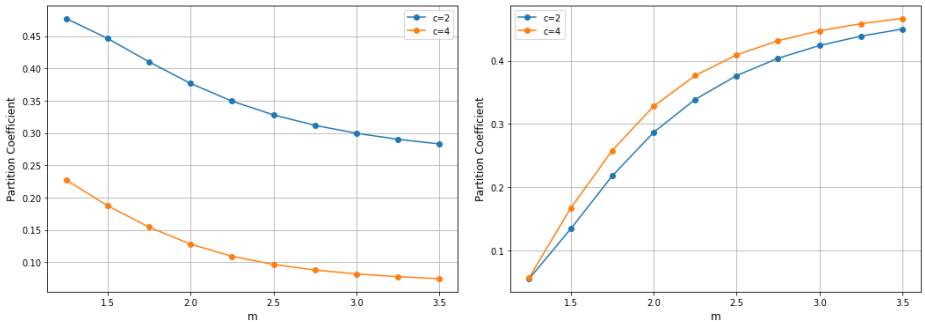

Figura 4. Coeficientes PC e PE obtidos para o conjunto de atributos formado por 3 componentes principais. Em azul, os valores são referentes ao experimento realizado utilizando 2 clusters. Em laranja, os valores representam os experimentos realizados com 4 clusters.

visual, nota-se que a clusterização obtida com e sem o PCA ficou próxima. O que foi observado foi uma troca das 'labels' pela função utilizada.

Resultados obtidos para este algoritmo são apresentados de forma mais detalhada na seção seguinte, avaliando a performance e os clusters resultantes para este em conjunto com o FCM.

6 Análise Comparativa

6.1 Conjunto de atributos Original

Os resultados obtimos para o FCM e para o K-Means utilizando 4 clusters e o conjunto de atributos original estão sumarizados na Figura 6. Os gráficos dos clusters propostos pelo FCM indica que poderia benéfico investigar o caso de lançar campanhas voltadas para o cluster 2, que são os clientes com um salário baixo e um gasto também baixo.

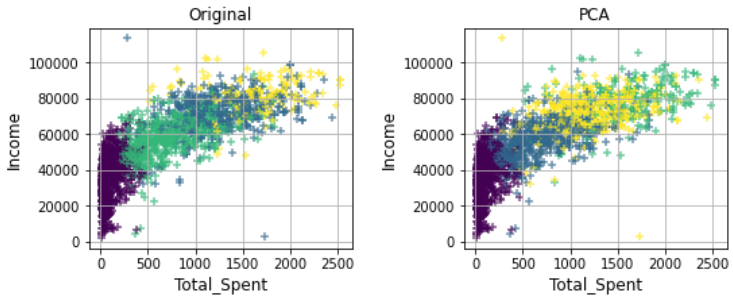

Figura 5. Clusters obtidos para os dois experimentos realizados com o K-Means.

Para o K-Means, analogamente ao que foi indicado pelo FCM, o cluster referente aos clientes com menor salário se refere também aos clientes que menos consomem produtos da empresa. Seria interessante pensar em campanhas para esse público alvo, realizando análises estatísticas mais específicas voltadas para o grupo. Alguns gráficos apresentando os clusters após realizada uma correspondência por inspeção dos grupos gerados estão apresentados no Anexo A.

6.2 Conjunto de atributos com PCA

Ao utilizar o PCA, os resultados obtidos para cada método apresenta maior divergência em relação aos clusters formados, e em relação ao número de amostras presente em cada cluster. Os resultados obtidos para ambos algoritmos estão reunidos na Figura 7.

No experimento com o conjunto de atributos original, o FCM havia indicado apenas 1 cluster com baixo consumo dos produtos da empresa – refente a clientes com salário de baixo a médio. Com o experimento utilizando dados de PCA, o modelo passa a indicar clusters mais "intuitivos": o cluster 0 abriga os clientes com maiores salários e logo apresenta o maior consumo. O cluster 2, que se posiciona no gráfico de dispersão um pouco abaixo e à esquerda do 0, tem um consumo mediano indicado no box-plot.

Adicionalmente, o FCM indica um cluster com baixo consumo indicado pelo número 3 de clientes que possuem salário de baixo a médio; e também um outro cluster com salário muito baixo e consumo muito baixo, indicado pelo número 1. Logo, o FCM parece ter se beneficiado da aplicação de PCA no conjunto de dados, resultando em um resultado com grupos mais distintos entre si.

Para o K-Means, o modelo continua apontando 3 clusters com consumo de médio a alto, e um cluster de consumo inferior – referente aos clientes com salário de baixo a médio. Porém, com o PCA, o K-Means aglomera mais clientes em seu cluster 0 (aproximadamente metade dos clientes do conjunto de dados). Este conjunto indicado é o de menor consumo quando comparado aos demais, o que é evidenciado ao analisar o box-plot dos clusters. Essa aglomeração maior em um único cluster pode indicar que a realização de uma campanha voltada para este

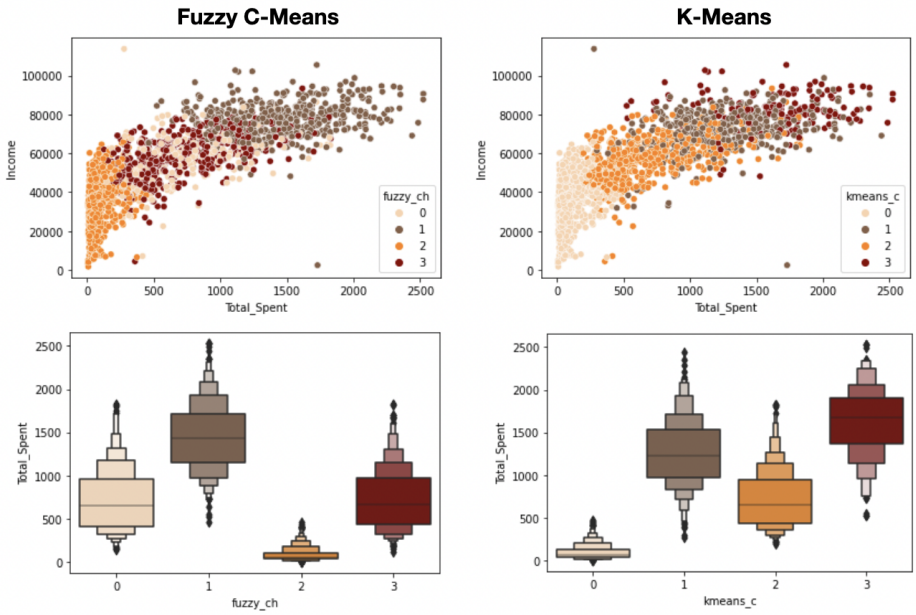

Figura 6. Variáveis Income e Total_Spent para os algoritmos Fuzzy C-Means e K-Means modelos utilizando o conjunto de atributos original.

grupo de clientes seja mais eficaz e mais simples de direcionar ao público-alvo do que o investimento em dois grupos distintos proposto pelo FCM com PCA.

Por outro lado, os resultados obtidos com o FCM parece apresentar uma maior granularidade com relação ao perfil do consumidor. O que pode ser avaliado futuramente são os grupos de consumidores segmentados testando diferentes valores do coeficiente nebuloso, como uma forma de ajuste da resposta do modelo à nebulosidade inerente ao conjunto.

7 Conclusões

O trabalho mostrou a viabilidade da aplicação do algoritmo Fuzzy C-Means em conjuntos de dados nebulosos, que apresentam uma separação em partições "fuzzy". Utilizando o conjunto de atributos original, os resultados obtidos para ambos algoritmos apresentaram certa proximidade, com clusters aparentemente "correspondentes"e com números de amostras parecidos. Ao aplicar o PCA nos atributos do conjunto de dados, os clusters propostos pelo FCM e pelo K-Means são mais distintos.

O FCM se mostrou mais sensível à dimensão do conjunto de dados, apresentando clusters mais distintos entre os experimentos com e sem PCA. Adicionalmente, o FCM parece ter apresentado uma maior sensibilidade com relação ao

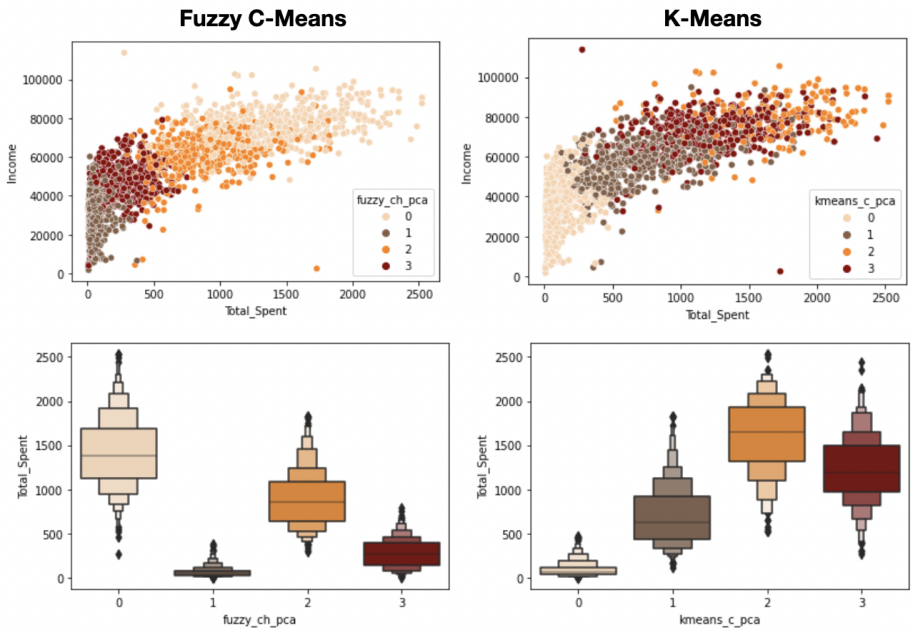

Figura 7. Variáveis Income e Total_Spent para os algoritmos Fuzzy C-Means e K-Means modelos utilizando o conjunto de atributos obtido aplicando PCA com 2 componentes principais.

perfil dos consumidores, sugerindo conjuntos mais "segmentativos" com relação às variáveis de interesse consideradas.

Referências

1. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers Geosciences **10**(2), 191–203 (1984). https://doi.org/https://doi.org/10.1016/0098-3004(84)90020-7, `https://www.sciencedirect.com/science/article/pii/0098300484900207`
2. Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J.: Fuzzy c-means clustering with spatial information for image segmentation. computerized medical imaging and graphics **30**(1), 9–15 (2006)
3. Hamerly, G., Elkan, C.: Learning the k in k-means. Advances in neural information processing systems **16**, 281–288 (2003)
4. Junior, A.C., Palhano, M.B., Felippe, G., de Menezes, C.T., de Azevedo Simões, P.W.T., de Mattos Garcia, M.C.: Os algoritmos fuzzy c-means, robust c-prototypes e unsupervised robust c-prototypes aplicados à uma base de dados das bacias hidrográficas da região de criciúma. Anais SULCOMP **6** (2013)
5. Patel, A.: Customer personality analysis (Aug 2021), `https://www.kaggle.com/imakash3011/customer-personality-analysis/metadata`

6. Rokach, L., Maimon, O.: Clustering methods. In: Data mining and knowledge discovery handbook, pp. 321–352. Springer (2005)
7. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017). https://doi.org/https://doi.org/10.1016/j.neucom.2017.06.053, https://www.sciencedirect.com/science/article/pii/S0925231217311815
8. Schwämmle, V., Jensen, O.N.: A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. Bioinformatics **26**(22), 2841–2848 (09 2010). https://doi.org/10.1093/bioinformatics/btq534, https://doi.org/10.1093/bioinformatics/btq534

A Anexo: Clusters obtidos com o conjunto de atributos original para algumas variáveis de interesse

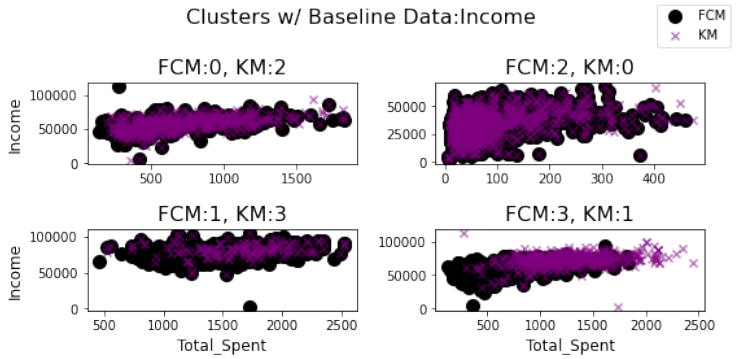

Figura 8. Clusters obtidos para o FCM e o K-Means para as variáveis "Total_Spent"e "Income".

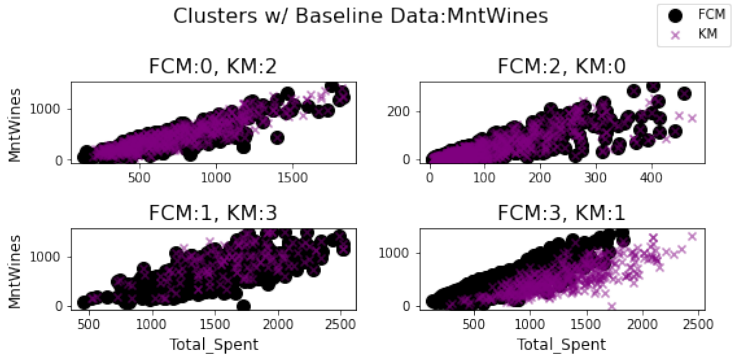

Figura 9. Clusters obtidos para o FCM e o K-Means para as variáveis "Total_Spent"e "MntWines".

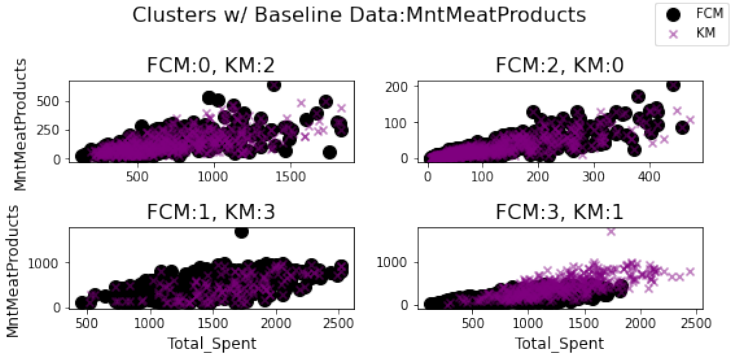

Figura 10. Clusters obtidos para o FCM e o K-Means para as variáveis "Total_Spent"e "MntMeatProducts".

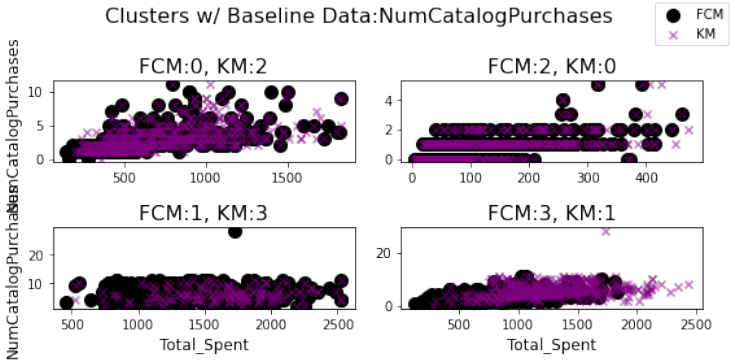

Figura 11. Clusters obtidos para o FCM e o K-Means para as variáveis "Total_Spent"e "NumCatalogPurchases".