Uso de Técnicas de Clusterização em uma Base de Dados Financeira

Use of Clustering Techniques in a Financial Database

Armando Pereira Pontes Júnior[1] orcid.org/0000-0002-8212-4589
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: appj@ecomp.poli.br

Resumo

O artigo tem como foco o uso de duas importantes técnicas computacionais para problemas de clusterização. Os algoritmos utilizados foram o *K-Means* e o *Fuzzy C-Means* (FCM), que aplicados em uma base de dados financeira de concessão de crédito pessoal podem auxiliar o tomador de decisão a identificar as principais características dos mutuários que se encontravam adimplentes e mutuários que estavam inadimplentes. O processo de clusterização investigou, através de 15 características (divididas entre características pessoais, condições de emprego e renda e condições da operação de crédito), similaridades que pudessem ajudar na formação de k grupos distintos. O resultado demonstra que as técnicas de agrupamentos aplicadas podem ser eficientes como ferramentas complementares para auxiliar o gestor financeiro nas suas atividades de classificação de risco, tomada de decisão e gerenciamento do crédito.

Palavras-Chave*:* Clusterização; Adimplência, Inadimplência; *K-Means*, *Fuzzy C-Means*.

Abstract

The article focuses on the use of two important computational techniques for clustering problems. The algorithms used were the K-Means and the Fuzzy C-Means (FCM), which have a financial database of credit granting, which help the decision maker to identify the main borrowing factors that were in arrears and borrowers that were defaulting. The clustering process investigated, through 15 characteristics (personal comparisons, income subsidies and operating conditions), similarities that can help in the formation of work groups. The following evidence that will be stored in companies in the activities of the date will be such as classifications to backup the risk and risk.

Key-words: *Clustering; Payment, Default; K-Means, Fuzzy C-Means.*

http://dx.doi.org/10.25286/repa.v3i3.976

1 Introdução

1.1 Intermediação Financeira

O nível de atividade econômica de um país é essencial para o crescimento, para a geração de emprego e renda e para a melhoria das condições socioeconômica de sua população. A atividade produtiva é altamente dependente, por um lado, dos investimentos realizados pelas empresas na produção de bens e serviços, e por outro lado, do consumo, que tem sua maior fatia realizada pelas famílias. Dados do IBGE - Instituto Brasileiro de Geografia e Estatística revelam que no ano de 2017 o consumo das famílias foi responsável por 63,43% do PIB, quando consideramos o cálculo do PIB pela renda **[1]**.

A ausência de capital, parcial ou total, dos agentes econômicos na demanda real ou latente, por consumo dos bens e serviços ofertados pelo mercado é facilmente verificado. É neste contexto de desequilíbrio entre os agentes superavitários e os agentes deficitários que surge a importante figura do crédito.

As instituições financeiras atuam como agentes de intermediação financeira no mercado, captando recursos juntos aos investidores pessoas físicas, empresas e Governos que possuem fundos excedentes e canalizam àqueles que necessitam de recursos para financiar seu déficit orçamentário **[2]**. A intermediação financeira é uma atividade que requer algumas condições básicas, tais como a existência de moeda, a consolidação de uma base legal e institucional e a existência de agentes econômicos superavitários e deficitários. Atendidas as duas primeiras condições, as instituições financeiras prestam o papel de intermediação entre os agentes econômicos, o primeiro, que possuem recursos financeiros em abundâncias e estão dispostos a emprestar, e o segundo, que necessitam de aportes extras para equilibrar suas finanças ou fazer frente a novos investimentos.

1.2 Risco de Crédito

Por definição, crédito é todo ato de vontade de alguém (pessoa física, jurídica ou Governo) em ceder, temporariamente, parte de seu patrimônio a um outro (pessoa física, jurídica ou Governo) com a expectativa de que essa parcela do patrimônio volte a sua posse de forma integral, acrescida de remuneração, e que seja feito no tempo aprazado **[3]**. De forma mais simples, crédito é o ato de entregar um certo valor mediante promessa de pagamento futuro de um montante maior ao que foi emprestado.

Entretanto, em toda operação de crédito é inerente a figura do risco, algo a ser administrado, mas nunca poderá ser completamente eliminado. A melhor forma de gerenciar a concessão de crédito é através da elaboração de mecanismos de identificação, mensuração e classificação de riscos, de uma forma que o tomador de decisão possa usufruir de ferramentas que o ajude a minimizar o risco de crédito.

Assim, este artigo investigará duas importantes técnicas de agrupamentos que podem servir de ferramentas complementares na tarefa de administração do risco de crédito.

Este artigo encontra-se dividido em seis seções. Além da introdução, a segunda seção trata de forma simples duas importantes técnicas de *clusterizaçã*o e detalha os dois algoritmos utilizados no artigo, mostrando a lógica de suas execuções. A terceira seção descreve em detalhes a base de dados financeira usada para o estudo de caso e como foi feito o pré-processamento. Na quarta seção são explicadas as métricas de validação usadas na execução dos algoritmos e também são apresentados os resultados numéricos. Na quinta seção é apresentada uma análise comparativa dos resultados. E por fim, a sexta e última seção discorre sobre as conclusões e as possíveis contribuições futuras.

1 Clusterização

2.1 Teoria

Os aumentos consideráveis na geração de dados demandam cada vez mais o uso técnicas que são capazes de realizar a extração do conhecimento de forma eficiente e automática.

Desta forma, podemos definir *clusterização* como processos computacionais muito utilizados em *Data Mining, e* que são bastante úteis na resolução de problemas de classificações e agrupamentos de conjunto de dados **[4]**.

A análise de clusters envolve a organização de um conjunto de padrões, comumente

representado na forma de vetores de atributos ou pontos em um espaço multidimensional, em grupos de acordo com uma medida de similaridade.

Existem diversas medidas de similaridade, a depender da natureza do problema que estamos a tratar. Assim, a formação dos grupos dependerá exclusivamente dos critérios de similaridade pré-definidos e da escolha das técnicas a serem utilizadas.

Neste artigo serão utilizadas duas técnicas bastante difundidas, *K-Means* e o *Fuzzy C-Means* (FCM). Depois serão comparadas as respectivas respostas de cada algoritmo para o problema de partição de uma base financeira que apresentam dados de bons (adimplentes) e maus (inadimplentes) pagadores.

2.2 *K-Means*

É um algoritmo do tipo não supervisionado proposto por MacQueen em 1967 **[5]**. Ele é muito eficiente e ao mesmo tempo de simples execução, o que o faz ser bastante utilizado para resolver o bem conhecido problema de clusterização. Em 2009, o *K-Means* foi considerado um dos dez algoritmos mais importante no campo da mineração de dados, considerando, como já mencionado, sua simplicidade mais também a sua escalabilidade. O *K-Means* possui complexidade ***O*** (t. N. K), onde t é a quantidade de iterações, N é a quantidade de objetos e K é a quantidades de grupos. Portanto, a complexidade é linear para qualquer variável do problema. Todavia, o algoritmo apresenta algumas restrições quando aplicado em bases mais complexas. Um dos problemas é fato da escolha inicial dos centroides (a lógica do algoritmo será melhor detalhada no próximo tópico) poder interferir nas soluções apresentadas. Isso o leva muitas vezes a obter soluções convergidas para ótimos locais.

2.3 Lógica do *K-Means*

O algoritmo se baseia na subdivisão de um conjunto de dados em k subgrupos, onde cada observação pertencerá apenas e, somente apenas, a um único subgrupo. Assim, dado um conjunto de observações $(x_1, x_2, \ldots, x_n)$ em que cada x_i tem d-dimensões (a dimensão representa a quantidade de características da observação x_i), o algoritmo *K-Means* particionará as n observações em k subgrupos ($2 \leq k \leq n$), que guardem o maior grau de semelhança entre suas observações.

Este artigo utilizará a distância euclidiana como a função de similaridade. Assim, dadas duas observações quaisquer, a distância euclidiana entre elas é calculada de acordo com a equação:

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

O passo seguinte é associar cada observação ao centroide mais próximo. Quando nenhuma observação estiver pendente, a primeira iteração estará concluída. O passo seguinte é recalcular a posição dos centroides tornando-os baricentros dos subgrupos definidos na etapa anterior. O cálculo leva em consideração as médias das distâncias euclidianas, novamente de acordo com a equação (1), das observações dentro de cada grupo. O recalculo da posição dos centroides é dado por:

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

Onde C_k representa o baricentro do grupo K e n_k é o total de observações associadas ao cluster K. A partir daí o algoritmo entra na sua regra de *loop*, até que um dos critérios de parada ocorra. Os critérios mais comuns para interromper as iterações são: a) não haja mais mudança na posição dos centroides e b) se atinja o número máximo de iterações predefinidas pelo usuário.

O Algoritmo 1 abaixo representa, de forma simplista, os passos para implementação do *K-Means* clássico, onde a base de dados com *i* observações, *d* dimensões e *K* centroides são os parâmetros de entrada do algoritmo.

http://dx.doi.org/10.25286/repa.v3i3.976

Algoritimo 1:	K-means Clássico.
Entrada:	Conjundo de daddos, K centroides.
Saída:	Base de dados dividida em K grupos.
1.	Escolher K centroide aleatoriamente.
2.	Calcular a distância de cada objeto aos centroides, conforme equação (1).
3.	Atribuir cada objeto ao seu centroide mais próximo.
4.	Atualiza a posição dos centroides, conforme equação (2), para as médias das instâncias de cada grupo.
5.	Repete passos 2, 3 e 4 até que nenhum centroide mude de posição ou se atinja o número máximo de iterações.
6.	FIM.

2.4 *Fuzzy C-Means* (FCM)

Alguns problemas envolvem grupos mais delineados que não podem ser separados de uma maneira *hard*, como é feito no *K-Means.* Em outras palavras, há situações em que as categorias se sobrepõem umas às outras e em diferentes níveis. Nesses casos pode-se recorrer a lógica *fuzzy*, ou seja, em alguns agrupamentos as observações pertencem a todos grupos, com diferentes graus ou níveis de pertinência, assumindo valores contínuos de pertinência (o que contrapõe a lógica binária do *K-Means*). Assim, tratando da possibilidade de partição com sobreposição (*overlapping,* em inglês) diversos algoritmos foram sendo apresentados, dentro os quais o *Fuzzy C-Means* (FCM) que foi introduzido em 1984 por Bezdek **[6]**.

2.5 Lógica do *Fuzzy C-Means* (FCM)

Quando um algoritmo fuzzy é aplicado a um conjunto de dados, o resultado é uma matriz *fuzzy* de modo que:

$$\begin{cases} P = [p_{i,j}] \\ p_{i,j} \in [0,1] \end{cases} \quad \textbf{(3)}$$

Onde a P é uma matriz de dimensão K x N, sendo que o *K* representa a quantidade de grupos e o *N* a quantidade de objetos. O valor de cada $p_{i,j}$ é o grau de associação do *j-ésimo* objeto ao *i-ésimo* grupo *fuzzy*. Assim, todos os objetos possuem algum grau de pertinência com todos os grupos, inclusive pertinência de valor nulo. Para se executar o algoritmo deve-se seguir as seguintes etapas:

1. Defina-se o número *c* de grupos *fuzzy*, com $2 \leq c \leq n$;
2. Defina-se o valor do coeficiente de "*fuzzificação" m*;
3. Inicializa-se a matriz de pertencimento $P^{(0)}$ com valores aleatórios;
4. Calcula-se os centroides de cada grupo *c*, com a seguinte equação:

$$c_{i,j} = \frac{\sum_{c=1}^{n}(p_{i,j})^m x_i}{\sum_{c=1}^{n}(p_{i,j})^m} \quad \textbf{(4)}$$

5. Calcula-se a distância euclidiana $D_{i,j}$, como mostrado na equação (1), entre cada ponto *i* para cada centroide *j*;
6. Atualiza-se os valores $p_{i,j}$ da matriz de pertencimento P, conforme a seguinte equação:

$$p_{i,j} = \left[\sum_{c=1}^{c}\left(\frac{D_{i,j}}{D_{c,j}}\right)^{\frac{2}{m-1}}\right]^{-1} \quad \textbf{(5)}$$

7. Volta-se a executar as etapas de 4 a 6 até que o módulo da diferença entre as duas matrizes de pertencimento, a atual e a da iteração anterior, seja menor que um coeficiente de erro εdefinido pelo usuário. Formalmente:
8.

$$\| P^t - P^{t-1} \| < \varepsilon \quad \textbf{(6)}$$

Onde o (*t*) representa a iteração atual, (*t-1*) a iteração imediatamente anterior.

3 Base de Dados

A base de dados apresentada em sua forma original contava com 40.320 registros de operações financeiras de crédito pessoal. Cada registro continha 21 características conforme apresentado na Tabela 1.

Tabela 1. Dicionário de dados (original)

ID	ID do mutuário
Setor de Atividade	Atividade que mutuário exerce
Data de Nascimento	Data de nascimento, dd/mm/aaaa
Data da proposta	Data da primeira porposta
Estado Civil	Estado civil
Sexo	Sexo
Grau de Instrução	Grau de Instrução
CEP	CEP residencial
UF_Endereço	UF residencial
DDD Residencial	Código de DDD residencial
Tipo Tel. Residencial	Tipo de telefone que possui na residência
Tipo de Residência	Situação e tipo da sua residência
Classe Profissional	Categoria profissional
Renda	Renda salarial do mutuário
Outras Rendas	Demais renda, exceto salarial
Dependentes	Número de dependentes
Renda do Cônjuge	Renda (diversas) do cônjuge
Produto Comercializado	Identificação do produto comercializado: cartão de crédito, CDC, cheque especial
Idade	Idade do mutuário (em anos)
Tempo de Emprego	Tempo no emprego (em anos)
Alvo	Identficação da situação do contrato: adimplência/inadimplência

3.2 Pré-processamento

Com objetivo de tratar, limpar, organizar e melhorar a apresentação dos dados foi feito um pré-processamento visando a remoção das observações onde verificamos total ausência de informação ou valores incoerentes com atributo. À título de exemplo, foram retiradas as observações que possuíam valores negativos para o atributo Renda ou valores acima de 120 (anos) para o atributo Idade.

Posteriormente foram feitas as remoções dos atributos ID, Setor de Atividade, Data de Nascimento, CEP, UF_Endereço, DDD Residencial e Tipo Tel. Residencial. O primeiro atributo removido foi o ID, que apenas identificava a quantidade de registros existente na base (40.320). Os atributos Setor de Atividade e CEP apresentavam uma gama diversa de informações que por si só não agregavam muito valor a base (particularmente em relação ao atributo Setor de Atividade, verificou-se que a mesma informação estava registrada com escritas diversas ou com uso de abreviações). Já o atributo Data de Nascimento foi removido uma vez que guardava redundância com o atributo Idade, sendo este último mais fácil de manusear. E de forma empírica, por entender que esses atributos não trariam ganhos de informação à pesquisa, foram feitas as remoções dos atributos CEP, UF_Endereço, DDD Residencial e Tipo Tel. Residencial.

O próximo passo foi categorizar cada atributo que se apresentava como *string*, dando as possíveis classificações do atributo um valor discreto. À título de exemplo, o atributo Sexo que podia ser classificado em "masculino" ou "feminino" foi reclassificado com os valores 1, para feminino e 2, para masculino. E ainda foi criada um atributo chamado de Idade2 que categorizava o valor discreto de cada idade dos mutuários em faixas etárias (até 25 anos, de 25 a 35 anos, de 35 a 45 anos, de 45 a 65 anos, acima de 65 anos).

Por fim, após todos os atributos já se apresentarem na forma numérica, foram feitas as normalizações para o intervalo de [0,1].

Ao concluir a fase de pré-processamento, a base de dados que foi submetida a execução dos algoritmos contava agora com 28.700 registros e 15 características.

4 Resultados

4.1 Métricas - Definições

Os resultados preliminares foram submetidos ao crivo das quatro métricas, no intuito de verificar a qualidade dos resultados e também o valor ideal de k. As observações foram feitas em cada uma das categorias do atributo Alvo (adimplentes e inadimplentes) e, também, por algoritmo executado. As métricas utilizadas foram as seguintes:

Estatística GAP: tem por objetivo encontrar um número ideal de *cluster*. Seu cálculo é feito pela diferença do logaritmo da distância *intra-cluster* do grupo analisado e de um conjunto de dados aleatório. Trata-se, portanto, de maximizar a diferença entre as distâncias do agrupamento

http://dx.doi.org/10.25286/repa.v3i3.976

escolhido e de um agrupamento aleatório. O objetivo é mostrar que o agrupamento escolhido é diferente de um aleatório.

Distância *Intra-Cluster*: é utilizada para calcular a distância de duas observações pertencentes ao mesmo cluster. Sua otimização se dar quando os valores são baixos, que indica proximidade das observações dentro do cluster.

Distância *Inter-Cluster*: é a métrica utilizada para calcular a distância entre dois centroides. Seu valor de otimização se dar quando os valores crescem, que demonstram que os centroides são realmente díspares.

Erro Quantizado: uma forma de avaliar a quantização do espaço obtido mediante a aplicação de um algoritmo de agrupamento é a lógica desta métrica. Ela está baseada no cálculo da média das distâncias entre os dados e o vetor que representa a região onde eles estão localizados. É uma métrica que avalia a eficiência do algoritmo para valores crescentes de K. Ela é otimizada quando se têm valores baixos.

4.2 Resultados das Métricas

As Tabelas 2 e 3 apresentam os resultados das simulações dos dois algoritmos (*K-Means* e FCM) para os dois grupos pesquisados (adimplentes e inadimplentes). Para que os resultados apresentassem consistência estatística, foram feitas 30 execuções para cada quantidade de *K* desejado nos dois algoritmos e por cada categoria do atributo Alvo (adimplentes e inadimplentes). Os valores escolhidos para K partiram de 2 até 10 *clusters*.

A escolha pelo número ideal de K dos grupos será feita observando primordialmente a estatística GAP, visto que essa métrica é bastante eficiente para escolha da melhor quantidade de grupos **[7]**.

Tabela 2. Resultados dos algoritmos utilizando 30 simulações e condição de parada 1.000 iterações . Resultados para o grupo de adimplentes

Algoritmo-K	GAP	Distância Intra-Cluster	Distância Inter-Custer	Erro Quantizado
K-Means - 2	**0,185 (0,063)**	15281,714 (146,79)	2,256(0,125)	0,873(0,009)
K-Means - 3	0,317(0,145)	13779,275 (486,18)	7,33(0,570)	0,77(0,011)
K-Means - 4	0,456(0,109)	12848,772 (439,14)	15,513(1,210)	0,72(0,023)
K-Means - 5	0,516(0,112)	12025,928(197,20)	27,809(1,764)	0,677(0,015)
K-Means - 6	0,560(0,132)	11617,746(265,07)	42,896(1,977)	0,658(0,013)
K-Means - 7	0,586(0,112)	11136,261(209,23)	60,838(2,879)	0,632(0,014)
K-Means - 8	0,634(0,130)	10710,921(144,57)	82,663(3,257)	0,612(0,011)
K-Means - 9	0,643(0,120)	10459,493(153,54)	109,637(3,555)	0,596(0,014)
K-Means - 10	0,690(0,107)	**10269,051(180,47)**	**138,55(6,133)**	**0,584(0,012)**
FCM - 2	0,182 (0,056)	16109,797(1,854)	0,298(0,001)	0,920(0,000)
FCM - 3	**0,409(0,061)**	14290,551(1,338)	2,308(0,306)	0,825(0,000)
FCM - 4	0,381(0,075)	14263,469(15,246)	4,017(0,039)	0,839(0,000)
FCM - 5	0,416(0,081)	14214,267(8,488)	6,891(0,471)	0,810(0,000)
FCM - 6	0,383(0,091)	14190,755(38,057)	9,327(0,341)	0,845(0,009)
FCM - 7	0,476(0,085)	13902,936(435,158)	14,793(2,408)	0,815(0,12)
FCM - 8	0,532(0,014)	13566,182(370,003)	21,394(2,876)	0,786(0,048)
FCM - 9	0,534(0,112)	13506,843(216,494)	26,372(3,343)	**0,775(0,028)**
FCM - 10	0,519(0,095)	**13453,251(180,276)**	**32,675(0,400)**	0,807(0,027)

Tabela 3. Resultados dos algoritmos utilizando 30 simulações e condição de parada 1.000 iterações . Resultados para o grupo de inadimplentes

Algoritmo-K	GAP	Distância Intra-Cluster	Distância Inter-Custer	Erro Quantizado
K-Means - 2	0,224(0,120)	8376,685(475,019)	1,978(0,160)	0,752(0,045)
K-Means - 3	**0,338(0,107)**	7661,666(89,506)	5,893(0,793)	0,695(0,034)
K-Means - 4	0,353(0,141)	7328,630(120,082)	12,012(1,039)	0,653(0,027)
K-Means - 5	0,499(0,105)	6907,897(117,411)	22,839(2,198)	0,632(0,019)
K-Means - 6	0,517(0,139)	6636,410(97,191)	35,001(3,300)	0,613(0,018)
K-Means - 7	0,591(0,131)	6385,989(77,171)	50,334(4,014)	0,594(0,016)
K-Means - 8	0,635(0,128)	6167,879(75,412)	70,889(4,865)	0,578(0,020)
K-Means - 9	0,708(0,116)	6002,152(107,063)	94,405(6,328)	0,567(0,014)
K-Means - 10	0,687(0,129)	**5842,265(90,158)**	**120,511(8,965)**	**0,549(0,013)**
FCM - 2	0,277(0,053)	8376,470(0,574)	0,952(0,001)	0,752(0,000)
FCM - 3	**0,346(0,057)**	8178,752(49,754)	1,955(0,066)	0,750(0,026)
FCM - 4	0,413(0,099)	7484,355(1,717)	5,730(0,009)	**0,673(0,000)**
FCM - 5	0,435(0,087)	7454,696(29,490)	8,488(0,156)	0,721(0,014)
FCM - 6	0,457(0,083)	7364,961(35,215)	12,593(0,903)	0,709(0,041)
FCM - 7	0,446(0,089)	7326,424(27,724)	17,193(0,117)	0,720(0,016)
FCM - 8	0,439(0,098)	7280,361(41,205)	22,866(0,754)	0,725(0,015)
FCM - 9	0,439(0,113)	7282,343(49,161)	28,452(0,169)	0,742(0,016)
FCM - 10	0,391(0,097)	**7249,279(52,416)**	**34,981(1,074)**	0,755(0,019)

5. Análise dos Resultados

5.1 Resultados do *K-Means* e do FCM Para o Grupo dos Adimplentes

Os resultados para o grupo de bons pagadores estão descritos na Tabela 2. Os valores se mostram consistentes com as métricas estabelecidas. Houve uma maximização da distância *inter-cluster* e uma minimização da distância *intra-cluster* nos dois algoritmos, à medida que o número de K aumentou. O erro quantizado também foi atendido, uma vez que houve diminuição desta métrica a medida que o número de cluster aumentava. Porém, houve uma pequena diferença, enquanto que no *K-Means* este valor se mostrou ótimo para K = 10, no FCM foi quando K atingiu o valor 9.

Com relação ao número de cluster ideal apontado principalmente pela métrica estatística GAP, o *K-Means*, maximizou no k=2 enquanto que o FCM foi no K=3.

Analisando uma amostra agrupada foi possível investigar de forma mais pormenorizada o perfil dos mutuários bons pagadores na criação, pelo K-Means, de dois *cluster*. Verificou-se que as características mais relevantes para formação do grupo foram "tempo de emprego", "prazo da proposta" e "produto contratado".

Já em relação ao FCM, como já descrito, o número ideal de cluster foi quando o K atingiu o valor 3. Verificando-se numa amostra agrupada que as características mais importantes para formação dos cluster foram "sexo", "produto contratado" e "prazo da proposta".

5.2 Resultados *K-Means* e do FCM Para o Grupo dos Inadimplentes

Os resultados para o grupo dos inadimplentes estão exibidos na Tabela 3. As métricas da distância *intra-cluster* e da distância *inter-cluster*

http://dx.doi.org/10.25286/repa.v3i3.976

obtiveram os resultados esperados, ou seja, minimizaram o valor da distância da primeira métrica e maximizaram o valor da distância da segunda. As duas métricas indicaram o k ideal como sendo de valor 10.

Também para a estatística GAP houve uma convergência dos dois algoritmos para o valor ideal de K = 3.

Agora analisando uma amostra agrupada do cluster k = 3 do *K-Means*, verificou que a principal característica para formação do cluster foi o "sexo" e o "tipo de residência".

Por fim, já em relação ao FCM, na análise mais apurada do agrupamento para K=3, verifica-se que as principais características para formação dos grupos foram "Tempo no Emprego", "Idade2" e "Estado Civil".

5.3 Análise Comparativa das Características dos Grupos Gerados Pelos Algoritmos

Adotou-se também uma outra abordagem comparativa para entender como cada algoritmo classificou os subgrupos do ponto de vista dos atributos. Assim, foram feitas as comparações de cada atributo estabelecendo sua importância para criação do subgrupo comparativamente com que foi estabelecido no outro algoritmo.

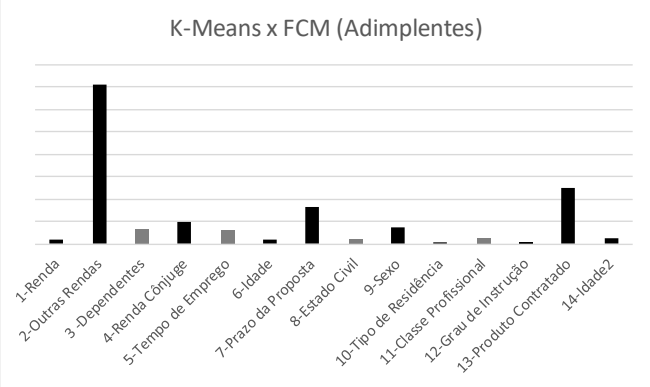

Figura 1: Comparativo das classificações.

A Figura 1 mostra como os algoritmos diferenciaram os 14 atributos da base de dados. Cada barra representa a diferença dada na importância dos atributos quando comparados *vis a vis* nos dois algoritmos. As barras na cor cinza, identificam que o FCM estabeleceu uma importância maior que o *K-Means* para um determinado atributo. As barras na cor preta dizem o contrário.

Esta mesma análise comparativa foi feita para o grupo de inadimplentes, como se pode observar na figura abaixo:

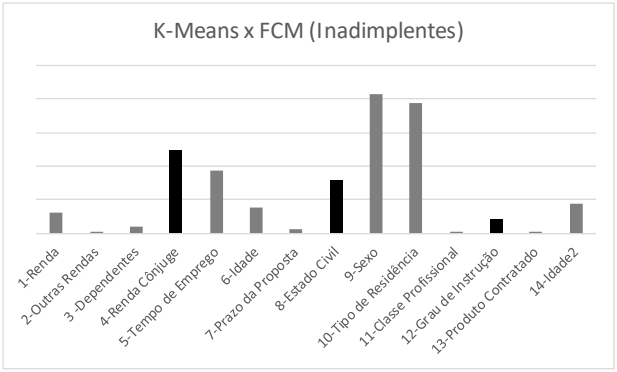

Figura 2: Comparativo das classificações.

Como se pode verificar, no grupo de adimplentes, os atributos que mais contribuíram para diferenciar o resultado do FCM do resultado do K-Means foram "Outras Rendas" e "Produto Comercializado". Enquanto que no grupo dos inadimplentes, os atributos "Sexo" e "Tipo de Residência" se mostraram como os mais relevantes para diferenciar os resultados obtidos em cada um dos algoritmos.

6 Conclusões

Este artigo analisou a aplicação de técnicas de *clusterização*, com a execução de dois importantes algoritmos: *K-Means* e *Fuzzy C-Means*. Foi utilizada uma base de dados financeira com informações sobre operações de crédito, dados pessoais e informações relativas à ocupação do mutuário. A base estava dividida em dois grupos, os adimplentes e os inadimplentes.

Assim, o objetivo era identificar perfis similares de mutuários adimplentes para que um gestor financeiro tivesse como estreitar a relação e potencializar os negócios. Bem como a importância de identificar perfis similares de mutuários inadimplentes para que o gestor fizesse uma administração mais próxima e cautelosa para com esse perfil de mutuário.

Do ponto de vista dos resultados apresentados pelos algoritmos, especificamente para o grupo das pessoas adimplentes, não foi possível compara-los. Enquanto que o *K-Means* retornou 2 como número ideal partição do conjunto de dados, o FCM fixou o número de cluster como

sendo 3. Contudo, ambos concluíram que as características "prazo da proposta" e "produto contratado" são bons rótulo para se fazer um agrupamento.

Já em relação ao grupo de inadimplentes, ambos os algoritmos chegaram ao número ideal de partição em 3 grupos. Porém, divergiram nas características que esses grupos têm que ser particionados. Podemos considerar que FCM mostrou uma performance levemente superior ao *K-Means*uma vez que primeiro apresentou um valor de estatística gap maior que a do segundo.

Como trabalhos futuros, poderíamos agregar esse trabalho na modelagem de um sistema de *credit score* como etapa de pré-processamento. Sistematizando que perfis de mutuários inadimplentes como encontrados neste estudo de caso pontuariam menos no sistema de *credit score*. Já os perfis de mutuários adimplentes teriam uma pontuação maior no sistema, uma vez que apresentam um risco menor de crédito.

Referências

[1] INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Quadros Completos: PIB 2017**. – IBGE. IBGE, 10 ABR. 2018. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-detalhe-de-midia.html?view=mediaibge&catid=2102&id=1800>. Acesso em 19.05.2018, 20:56:15.

[2] GITMAN, Lawrence J. **Princípios de Administração Financeira**. 7 ed. São Paulo: Habra, 2002.

[3] SCHICKEL, Wolfgang K. A**nálise de Crédito Concessão e gerência de empréstimos**. 5 ed. São Paulo: Atlas, 2000.

[4] ALAM, Shafiq et al. Research on particle swarm optimization based clustering: a systematic review of literature and techniques. **Swarm and Evolutionary Computation**, v. 17, p. 1-13, 2014

[5] MACQUEEN, James et al. Some methods for classification and analysis of multivariate observations. In: **Berkeley symposium on mathematical statistics and probability, 5**., 1967, Berkely. **Proceedings...** Berkely: University of California Press, 1967. p. 281-297.

[6] BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191-203, 1984.

[7] TIBSHIRANI, Robert; WALTHER, Guenther; HASTIE, Trevor. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411-423, 2001.

http://dx.doi.org/10.25286/repa.v3i3.976