ESTUDO SOBRE CLUSTERIZAÇÃO GEOGRÁFICA EM OPERAÇÕES DE DISTRIBUIÇÃO DE CARGA

Victor Gonzaga Generato (IC) e Prof. Dr. Arnaldo R. de Aguiar Vallim Filho (Orientador)

Apoio:PIVIC Mackenzie

RESUMO

Este projeto buscou o combinar técnicas de aprendizado de máquina com metaheurísticas para solução de um problema de “Clusterização”, que é um problema combinatório clássico da área de Mineração de Dados (Silva et al., 2016). No caso específico deste artigo o trabalho contemplou, de forma particular a área de logística, o agrupamento que representa um importante problema prático dessa área. O problema de clusterização trata de distribuição de exemplares de um conjunto de dados em grupos com características similares entre si. O estudo fez uso de bases de dados disponíveis para desenvolver experimentos, com características similares a situações reais no ambiente de logística, dentre elas estão as coordenadas dos exemplares, identificação dos mesmo e uma variável representando carga máxima de cada um. Foram utilizadas três técnicas de clusterização combinadas e baseadas na metaheurística *simulated annealing*, em aprendizado de máquina, com o uso do algoritmo *k-means* e envolvendo uma rede neural artificial. A combinação de técnicas permitiu o estabelecimento de parâmetros de comparação, além de fornecer algoritmos capazes de realizar distribuições de exemplares em clusters gerando resultados otimizados. Nos experimentos desenvolvidos os resultados se mostraram satisfatórios em patamares elevados com nível de performance próximos de 100%, gerando clusters de qualidade razoavelmente elevada.

Palavras-chave: *Clusterização*, Aprendizado de máquina, *Metaheurísticas*.

ABSTRACT

This project sought to combine machine learning techniques with metaheuristics to solve a Clustering problem, which is a classic combinatorial problem in the Data Mining area (Silva et al., 2016). In the specific case of this paper, the work included, in a particular way, the logistics area, the grouping that represents an important practical problem in this area. The clustering problem deals with the distribution of copies of a data set in groups with similar characteristics. The study made use of available databases to develop experiments, with characteristics similar to real situations in the logistics environment, among which are the coordinates of the specimens, their identification and a variable representing the maximum load of each one.

Three combined clustering techniques based on simulated annealing metaheuristics were used, in machine learning, using the k-means algorithm and involving an artificial neural network. The combination of techniques allowed the establishment of comparison parameters, in addition to providing algorithms capable of performing specimen distributions in clusters generating optimized results. In the experiments developed, the results were satisfactory at high levels with a performance level close to 100%, generating clusters of reasonably high quality.

Keywords: Clustering, Machine Learning, Metaheuristics

1. INTRODUÇÃO

Este artigo trabalha com o problema de “clusterização” aplicado a uma operação logística. Clusterizar ou agrupamento é a tarefa de gerar grupos em uma base de dados de modo que os exemplares ou registros fiquem coerentes e semelhantes dentro de cada grupo. Simplificando, pode-se dizer que o objetivo de tal tarefa é gerar clusters com elementos parecidos entre si. Tratando-se da área de logística, essa tarefa diz respeito a pontos geográficos, definidos por suas coordenadas, que devem ter atendida sua demanda por serviços logísticos. Clusterizar pontos de demanda podem ser difícil e complexa devido a massiva fonte de dados que permite uma multiplicidade quase infinita de possibilidades de agrupamento gerando uma grande complexidade computacional para comparação entre as informações propostas. Este é assim, um problema de alta complexidade, estando na classe NP-Difícil, tornando inviável a solução por modelos exatos para problemas de grande dimensão. Com isso surge a necessidade de se criar algoritmos heurísticos que possibilitem esse trabalho e que sejam definitivamente eficazes e eficientes. O objetivo deste artigo é trabalhar com uma proposta de um modelo híbrido que combina técnicas de aprendizado de máquina com uma metaheurística clássica para solucionar o problema, de forma a unir e comparar esses algoritmos, procurando demonstrar suas efetividades.

1.1 PROBLEMA DA PESQUISA

O problema de clusterização, do ponto de vista prático, ocorre em muitos segmentos, como: agricultura, produção industrial, gestão empresarial, logística e outros. Uma grande parte desses casos é representada por modelos de grafos, sendo o problema resumido a agrupar nós de um grafo ou ainda, de particionamento de um grafo, uma importante aplicação da área (Glover, 1990; Pirlot, 1996). De modo mais formal Matteucci (2006) coloca que clusterização é o processo de se organizar elementos em grupos cujos membros são similares entre si, por algum critério.

No caso de operações logísticas de grande porte o número de pontos de demanda (pontos a atender) pode chegar aos milhares, e assim, o tratamento individual de cada ponto só é realmente feito nos processos operacionais finais quando se tem a definição final do atendimento daquele ponto específico. Porém, quando se trata de planejamento, os pontos (clientes a atender) são agrupados em clusters, pois, inclusive permitem que se tenha assim, uma visão mais apurada de tendências e padrões de comportamento do conjunto. A construção desses clusters constitui-se, assim, em etapa fundamental do planejamento logístico. Apenas como observação, deve-se salientar que o termo *districting*, é empregado em Logística para designar o processo de clusterização.

Diante deste quadro, nota-se a importância que um processo de clusterização pode exercer em operações logísticas, e como estudos de novas técnicas e algoritmos podem representar contribuições importantes para o conhecimento dessa área.

Considerando-se este contexto logístico dos problemas de clusterização, a questão que se coloca na presente pesquisa é:

"Como se comportaria a configuração de clusters de pontos de demanda de uma operação logística gerados por diferentes algoritmos? Em particular, se forem combinados métodos clássicos de clusterização com novas heurísticas e com soluções baseadas no uso de Inteligência Artificial?".

Espera-se com as respostas a estas questões não só agregar conhecimento novo sobre o comportamento dessas técnicas e do problema, mas, também, poder propor novos modelos e algoritmos para solução do problema.

Neste sentido, assim, vê-se que este é um projeto que pode trazer resultados relevantes do ponto de vista prático e científico.

1.2 OBJETIVO

Dado esse quadro, pode-se dizer que o objetivo geral deste projeto foi o desenvolvimento de um conjunto de experimentos para estudar a aplicação de diferentes técnicas combinadas no problema de "clusterização" de pontos de demanda de uma operação logística em uma dada área geográfica. A ideia foi comparar técnicas a fim de se identificar pontos fortes e fracos das abordagens.

No que se refere aos objetivos específicos da pesquisa, tem-se:

- Teste de diferentes técnicas de clusterização:

    . um método clássico de clusterização;

    . um processo heurístico de clusterização;

    . uma abordagem por Redes Neurais Artificiais.

- Desenvolver critérios para comparação de resultados entre as técnicas;

- Realizar experimentos com as diferentes abordagens, para diferentes instâncias, desenvolvendo-se comparações de resultados;

- A partir dessas análises, gerar uma base com os dados originais e com resultados dos experimentos, para que possam ser utilizadas para estudos exploratórios futuros, com o

intuito de abrir a possibilidade da descoberta de padrões e informações sobre esses processos.

2. REFERENCIAL TEÓRICO

Este capítulo tem por finalidade apresentar trabalhos científicos que tenham relação com o tema do presente projeto. A revisão dessas obras auxiliou no entendimento do problema e no conhecimento das estratégias e técnicas de solução com o objetivo de tentar tratar a questão de clusterização.

Segundo Ochi et al. (2005), o Problema de Clusterização (PC) consiste em dada uma base de dados X, agrupar (clusterizar) os objetos (elementos) de X de modo que objetos mais similares entre si sejam alocados ao mesmo cluster, enquanto que, objetos menos similares sejam alocados em outros clusters.

A questão da clusterização, na verdade, é um problema antigo, porém mais recentemente com o crescimento das aplicações de computação o assunto ganhou um novo impulso, em particular, com o crescimento da área de Mineração de Dados.

Em termos de abordagens, pode-se ter o problema de K-Clusterização com K pré-definido. Já quando K não é previamente conhecido, tem-se o Problema de Clusterização Automática (PCA). O que caracteriza este tipo de problema em mineração de dados, é que nesta área a base de dados costuma ter um porte elevado e o número de atributos de cada elemento é, em geral, alto (Ochi et al., 2005; Oliveira, 2005).

Já foi colocado anteriormente, que o problema de montagem de clusters se encontra na mais alta classe de complexidade, a NP-Difícil (Daskin, 1995). Desta forma, procurar soluções para um problema de grande porte, desenvolvendo uma pesquisa em todo o espaço de soluções é operacionalmente inviável, uma vez que o número de número de possibilidades cresce muito. Desta forma, as abordagens de solução não focam em métodos exatos, mas sim, em métodos heurísticos, incluindo-se aí as metaheurísticas.

Considerando-se os principais métodos, estes podem se subdividir em tipo hierárquico ou tipo de particionamento. O hierárquico, por sua vez, pode ainda, ser baseado em uma abordagem de aglomeração ou de divisão. A estratégia hierárquica de aglomeração é chamada de bottom-up, porque trabalha de baixo para cima, da base de todos os elementos individuais e vai-se agregando-os em clusters. O início deste processo assim, pode ser dar com cada elemento se constituindo em um cluster. Neste caso, cada cluster com tamanho maior que 1 terá sido composto por um conjunto de outros clusters. A união de clusters sempre ocorre com base em algum critério que leva em conta algum tipo de similaridade entre os

elementos da base de dados. Na outra abordagem hierárquica, aquela de divisão, é chamada de (top-down), já que trabalha de cima para baixo. Neste caso, o tem início com um único cluster que contém todos os elementos do conjunto de dados. Procede-se em seguida, a uma sequência de divisões, com base no critério que foi definido (Ochi et al., 2005).

A outra estratégia mencionada é a de particionamento. Neste caso, já se inicia o processo com um certo número “p” de clusters já definidos. Assim a partir desse p clusters, procede-se a uma troca de elementos entre esses clusters. Essas trocas são consideradas rearranjos do conjunto de clusters, pois a cada rearranjo tem-se uma nova configuração de clusters.

As configurações de clusters podem ser avaliadas por algum tipo de critério, que são denominados de função objetivo. Esta função tem sua importância, pois é fator decisivo do processo, devendo ser efetivamente representativa do objetivo que se busca alcançar. No que tange à área Logística as estratégias hierárquica e de particionamento têm sido usadas com frequência (D’Amico et al., 2002; Zhou et al., 2002; Vallim Filho, 2004).

Há vários trabalhos sobre o uso de métodos heurísticos ou de metaheurísticas aplicados a problemas de clusterização em Logística.

Zhou et al. (2002) geram zonas de atendimento de centros de distribuição, e isto é feito por uma metaheurística, o algoritmo genético. A abordagem tem por objetivo, balancear o custo de transporte e o nível de serviço.

D’Amico et al. (2002) apresentam uma revisão completa das jurisdições policiais na cidade de Buffalo, New York. O trabalho foi feito com a metaheurística simulated annealing, que é uma técnica que parte de uma solução inicial, e busca novas soluções de forma a minimizar a heterogeneidade entre as cargas de trabalho dos policiais.

Vallim Filho e Gualda (2003) apresentaram uma proposta de heurística de clusterização de pontos de demanda que deveriam ser atendidos por centros de distribuição de carga. Dentro desses clusters seriam identificadas posições que se constituíssem em “melhores candidatos” a centros de distribuição.

Vallim Filho e Gualda (2004) trataram a localização de instalações a partir de clusters com o uso de um modelo de otimização combinado à metaheurística *simulated annealing*. Vallim Filho (2004) faz uso de *simulated annealing*, para geração dos clusters e identifica a “melhor posição teórica” (MPT) em cada cluster para a localização de um centro de distribuição.

Novaes (1989) trabalhou clusterização em distribuição urbana, em que uma heurística subdivide uma área de estudo criando anéis a partir de quadrículas elementares. Novaes e

Graciolli (1999) propuseram um novo procedimento para criação de clusters em uma operação de distribuição. Neste caso, a região é subdividida em setores, que partem de um ponto central. Novaes et al. (2000) desenvolvem outro trabalho nessa linha, gerando distritos homogêneos adotando como critério uma medida do esforço de distribuição. O particionamento é feito por uma heurística e esta solução obtida é melhorada por meio de um algoritmo genético associado a um método de otimização direta, o método do gradiente.

Uma metaheurística muito conhecida que é utilizada em clusterização é a Busca Tabu ou *Tabu Search* (TS), proposta por Glover (1986).

Esta técnica foi inicialmente utilizada em problemas de programação matemática, mas depois disso, se expandiu para inúmeras áreas de atividade.

O problema de roteirização de veículos, em que cada rota representa um cluster, é um tipo de problema logístico que tem sido resolvido por *tabu search* (Crainic e Laporte, 1998). O problema clássico de partição de grafos também tem sido resolvido por esta técnica (Pirlot, 1996; Glover, 1990). E outro exemplo é o caso de células de manufatura, que também podem ser definidas por clusterização, e que têm sido resolvidas por TS.

Uma outra abordagem é apresentada por Furuta et. al (2005), que fazem uso de uma técnica de geometria computacional, o diagrama de Voronoi. Esta técnica foi aplicada a problema de montagem de distritos para alocação de ambulâncias.

Galvão et. Al (2006) também apresentaram um problema de districting em que fazem uso desse mesmo diagrama de Voronoi. Nesta aplicação, partem de um conjunto de clusters na forma de anéis, que por sua vez são subdivididos em setores. O processo vai relaxando as fronteiras iniciais dos clusters de forma iterativa até que uma convergência seja atingida.

Neste levantamento bibliográfico nota-se que há bastante literatura envolvendo métodos heurísticos aplicados à clusterização. A presente pesquisa procura trabalhar nessa mesma linha, de forma que a trazer uma contribuição a esse campo do conhecimento.

3. METODOLOGIA

A pesquisa deste projeto é do tipo “aplicada”, e adotou uma abordagem quantitativa. Sua finalidade é metodológica, já que, tem o intuito de propor novas abordagens estudando modelos e algoritmos de clusterização.

Em termos metodológicos, para início de projeto, foi necessário realizar um levantamento de bases de dados que serviriam de instrumento de estudo e meio de aplicação para os métodos e algoritmos que viriam a ser desenvolvidos. Com isso foram utilizadas algumas bases disponíveis no conhecido *dataset* de Uchoa et al. (2014). A escolha dessas

bases se deu por possuírem características de coordenadas bidimensionais, ou seja, o eixo X e o eixo Y, identificações dos pontos enumerados como se fossem “ID’s”, podendo assim fazer referências aos centros de distribuição de alguma indústria ou empresa e valores que poderiam ser considerados variáveis de peso abrindo a possibilidade de estudar um agrupamento com taxas de cargas máximas por cluster, como é corriqueiro na área de logística. Para analisar o comportamento dos algoritmos em função do tamanho dos dados, foram separadas cinco bases contando com uma estrutura de quatro colunas sendo elas “Id”, “Eixo X”, “Eixo Y” e “Peso a ser distribuído” e com a quantidade de 200, 401, 613, 800 e 1001 registros.

Os experimentos foram trabalhados com três tipos de técnicas aplicadas às mesmas instâncias de um problema de clusterização, são eles:

- um método clássico de clusterização;

- uma técnica de Inteligência Artificial, baseada em Redes Neurais Artificiais;

- uma metaheurística de clusterização, proposta por Vallim Filho (2004).

Tratando-se de um método clássico, decidiu-se estudar dois algoritmos de clusterização comuns no mercado. O Primeiro deles é o DBScan, proposto por Ester et al. (1996). O algoritmo é caracterizado por guiar o processo de descoberta de grupos com base na densidade de exemplares existentes na vizinhança de um exemplar já pertencente ao grupo. Esse por sua vez foi descartado pois trabalha com ruídos na clusterização, isto é, o algoritmo deixa pontos sem classificação de grupo podendo gerar outliers, isso diverge do resultado desejado já que para o estudo, não é possível, por exemplo, deixar centros de distribuição sem clusters. O segundo algoritmo estudado e por sua vez escolhido para aplicação na pesquisa foi o K-means. Sendo o mais popular entre os métodos clássicos, esse algoritmo inicia a classificação sorteando k pontos para serem os chamados centroides, sendo k um parâmetro de entrada do algoritmo e o número desejado de clusters. A partir do sorteio aleatório dos centroides, o processo de classificação começa a buscar por pontos parecidos levando em consideração as suas coordenadas, podendo assim dizer que em coordenadas X e Y, leva-se em consideração sua proximidade gerando partições com características semelhantes entre si.

A técnica de inteligência artificial baseada em redes neurais artificiais escolhida foi a também conhecida *Self-Organizing Maps* (SOM), ou Mapas de Kohonen, desenvolvida por Kohonen (1982). Esta técnica consiste em projetar os exemplares de um conjunto de dados em um espaço de baixa dimensão mantendo as relações topológicas dos pontos de forma que, sempre que for possível, os exemplares semelhantes fiquem próximos. O método de aprendizagem utilizado por esse algoritmo é o *competitive learning*, este por sua vez é uma

forma de aprendizagem não supervisionada onde os nós competem entre si para garantir o direito de representar um subconjunto de entrada. O resultado obtido do SOM é uma matriz unidimensional reunindo as projeções em grupos de características semelhantes, gerando assim os clusters desejados. Para o estudo em específico, a matriz de resultado foi construída com apenas uma coluna e com n linhas, sendo n um parâmetro de entrada e o número de clusters desejado. A utilização dessa técnica é realizada a partir da instalação da biblioteca "kohonen" na linguagem R, e do uso das funções e variáveis presentes na biblioteca.

Inicialmente foi estudado uma maneira de executar os métodos de clusterização K-Means e SOM por inúmeras vezes através de laços de repetição. Isso permite com que se obtenha várias soluções, diferentes distribuições de agrupamento entre os pontos e variados resultados para a função objetivo, ou apenas FO que foi definida conforme a equação 1:

$$\mathbf{FO = \sigma / \overline{x}} \quad (1)$$

onde: σ = o desvio padrão do volume de carga alocada a cada cluster;

$\overline{x}$ = a média do volume de carga alocada aos clusters.

A função objetivo definida dessa forma expressa o coeficiente estatístico de variação da carga alocada aos clusters. E o objetivo da heurística é a minimização desse coeficiente. Busca-se, portanto, clusters homogêneos em termos de volume de carga. Este é um objetivo importante em uma operação logística, pois torna homogêneo o esforço de produção entre os clusters, facilitando o planejamento operacional e reduzindo custos.

A FO é utilizada assim, para definir se a distribuição dos pontos de coordenadas em cada cluster é efetiva ou não. Com ela é possível analisar a proporção do desvio padrão da distribuição dos valores da variável tida como fator limitante do cluster, no caso do estudo é a variável peso, pela média dessa variável em cada grupo. Em termos de análise destes números quanto menor for o resultado de FO, mais bem distribuídos os pesos estão em cada cluster e assim por dizer, mais efetivo é o método de clusterização utilizado.

Para respeitar o limite máximo proposto por uma variável limitante, é realizado uma comparação entre o parâmetro de entrada (valor da variável limitante) e a somatória do campo analisado dentro de cada cluster. Foi desenvolvido uma inteligência que aumenta em um o número de clusters caso o resultado da comparação exceda o limite proposto. Isso permite que o algoritmo traga apenas resultados tocantes ao desejado e anula a necessidade de executar o algoritmo novamente mudando os parâmetros de entrada.

Sobre a última técnica utilizada no projeto, o *Simulated Anealling* foi utilizado como metaheurística de clusterização. O algoritmo que se fundamenta numa analogia com a termodinâmica simulando o arrefecimento de metais, foi usado para solucionar problemas de

análise combinatória nos clusters já segmentados pelas outras técnicas mencionadas. Originalmente pensado por Metropolis em meados de 1950, a 1ª proposta tinha como ideia gerar um modelo de processo de cristalização dos metais. Foi então somente em 1980 que um grupo de pesquisa independente notou as semelhanças do processo físico de recozimento (o *anealling*) com os problemas de análise combinatória. O processo do algoritmo é iniciado com uma temperatura T e um conjunto inicial de dados e conforme a temperatura abaixa, os exemplares dos conjuntos tendem a ficar mais bem organizados. A grande orquestradora dessas mudanças nos conjuntos é probabilidade de Boltzman conforme a equação 2:

$$B = \exp(-\Delta E/T) \quad (2)$$

onde: ΔE = diferença de energia (função objetivo) entre a solução atual e a solução do rearranjo

T = temperatura atual

Quando B é comparada a um número aleatório gerado entre 0 e 1, é utilizada para definir se um resultado pior que o atual pode ou não ser aceito para dar continuidade ao algoritmo, caso a solução seja inicialmente melhor que a atual a troca é feita automaticamente gerando assim um rearranjo no conjunto de dados. Ele termina quando atinge um limite mínimo da temperatura ou um número máximo de aceites de mudanças no conjunto.

No heurística estudada são utilizadas simultaneamente duas estratégias avaliativas: diversificação e intensificação. No início, quando a temperatura está elevada, o algoritmo diversifica as soluções de modo que uma solução ruim possa levar subsequentemente a uma solução melhor que as anteriores. Já no fim da execução, quando a temperatura está menor, o algoritmo tende a se concentrar em torno das melhores soluções (intensificação).

A técnica utilizada para a troca efetiva de exemplares entre os clusters é chamada de rearranjo. Nela são analisados possíveis alterações a fim de melhorar a distribuição dos exemplares e consequentemente a solução ideal de clusterização. Este processo deve ser realizado com muita cautela para não desconfigurar os grupos e assim perder as características de cluster, pois pontos espalhados em coordenadas de forma não agrupada não trazem o resultado ou a análise desejada.

Um grande apoio no desenvolvimento dos estudos foi o uso de um algoritmo (uma biblioteca da linguagem R) que gera o chamado *convex hull*, ou “envelope convexo”. O *convex hull* de um conjunto de N pontos é definido como o menor subconjunto convexo desses pontos que contém dentro do seu perímetro todos os pontos do conjunto. No plano, é um polígono convexo de no máximo N lados (Brown, 1979; Jayaram e Fleyeh, 2016; Alshamrani, 2020). É, portanto, um algoritmo que permite descobrir as extremidades dos clusters num plano bidimensional. Com o *convex hull* tem-se assim o conjunto de pontos “extremos” de cada cluster. E foi esse o conjunto utilizado para se proceder aos rearranjos de pontos entre os

clusters. Com o uso do *convex hull*, além do tempo de execução diminuir consideravelmente, se tornou possível aplicar alterações de conjuntos somente nos pontos externos dos grupos reduzindo as chances de gerar resultados desordenados levando em consideração os limites de pesos por clusters. A descoberta dos pontos extremos de cada grupo trouxe à tona dois métodos de rearranjos usados no *Simulated Anealling*, são eles:

- Exemplar mais próximo;

- Exemplares presentes na circunferência de um exemplar escolhido.

No rearranjo utilizando o exemplar mais próximo é sorteado um ponto presente na lista obtida pelo *Convex Hull* e calculado a distância referente aos outros pontos, escolhendo assim o exemplar de menor distância para efetuar o rearranjo.

Já no rearranjo utilizando exemplares presentes na circunferência, é gerado um raio a partir das médias de distâncias entre os pontos resultantes do Convex Hull. Em seguida sorteia-se um dos pontos do conjunto de exemplares (Ponto 1) extremos do Convex Hull. A partir deste exemplar sorteado, desenha-se uma circunferência de raio igual à média de distância calculada. Identifica-se todos os pontos presentes nessa circunferência não pertencentes ao cluster em análise. Dentre esses pontos troca-se com o Ponto 1 aquele que apresentar a melhor solução.

Com a utilização do *Simulated Anealling* junto ao algoritmo do *Convex Hull* foi visto que não era necessário a execução dos métodos de clusterização iniciais em laços de repetição, pois eles já trariam a melhor solução e em um menor tempo de execução, mesmo partindo de uma solução distante da ”melhor”, ganhando mais performance no processamento.

O resultado disso gerou duas funções combinando um dos dois métodos de clusterização inicial com a metaheurística de otimização combinatória, formando funções do tipo KMeans + Simulated Anealling e SOM + Simulated Anealling.

4. RESULTADOS E DISCUSSÃO

Para a aplicação da metodologia descrita na seção anterior, todo o desenvolvimento foi feito na linguagem R, com uma parte sendo codificada e outra parte sendo utilizadas as suas bibliotecas.

As funções apresentadas na seção de metodologia acrescentaram à base de dados de entrada, um campo adicional contendo a indicação do cluster ao qual cada ponto foi alocado. Também é possível tirar informações de processamento como o tempo de execução e o decréscimo da função objetivo além de poder criar um gráfico bidimensional que

representa as distribuições dos clusters em coordenadas, onde cada cluster é representado por uma cor e os exemplares são coloridos em função disso.

Com a separação dos dois métodos de clusterização inicial foi possível gerar parâmetros de comparação afim de descobrir qual é o mais efetivo em termos de função objetivo, o mais rápido em tempo de processamento, ou até qual apresenta a melhor distribuição sem perder as características de um cluster. Vale ressaltar que para cada função foi aplicada a execução da metaheuristica Simulated Annealing para garantir a otimização combinatória.

As execuções nas bases de dados obtidas do *dataset* de Uchoa geraram os resultados presentes na tabela 1:

Tabela 1 - Resultados dos Experimentos

Base	No. de Registros	Método	FO Inicial	FO Final	% de Homogeneidade	Tempo de Execução (min)
uchoa200	200	Kmeans	0,3263157	0,080760	**91,92%**	3
uchoa200	200	SOM	0,3167526	0,068688	**93,13%**	4
uchoa400	401	Kmeans	0,3378948	0,015745	**98,43%**	4
uchoa400	401	SOM	0,4749167	0,012076	**98,79%**	4
uchoa600	613	Kmeans	0,2093854	0,001666	**99,83%**	7
uchoa600	613	SOM	0,2152622	0,005279	**99,47%**	5
uchoa800	801	Kmeans	0,1482959	0,006614	**99,34%**	6
uchoa800	801	SOM	0,2159741	0,006614	**99,34%**	7
uchoa1000	1001	Kmeans	0,291640	0,003549	**99,65%**	7
uchoa1000	1001	SOM	0,187917	0,001601	**99,84%**	6

A partir da tabela de resultados é possível analisar que a porcentagem de homogeneidade final é diretamente proporcional a quantidade de registros da base de dados. O fato de a base possuir uma quantidade maior de exemplares faz com que o número de combinações seja maior, podendo gerar um número maior de melhores soluções.

As funções se mostraram efetivas e plausíveis tratando-se de processamento e tempo de execução. É perceptível que em bases menores como a uchoa200, a execução dura em média três minutos e em bases maiores como a uchoa1000 o tempo dura em média sete minutos. Não há diferença significativa de tempo de execução entre os métodos *KMeans* e *SOM*, já que o que dita o processamento na verdade é o algoritimo de otimização combinatória aplicado *Simulated Anelling*.

Também é possível observar que o cálculo da função objetivo que de início variava em torno de 50% a 80% de homogeneidade na distribuição dos pesos por cluster, ao final variam

em torno de 91% a 99% de homogeneidade, o que em termos matemáticos é um resultado razoavelmente adequado. O aumento da homogeneidade está relacionado ao decréscimo da função objetivo, que ocorre ao decorrer da execução das funções. Essa diminuição do valor da FO pode ser observada quando colocada num gráfico que possui como eixo vertical a função objetivo e como eixo horizontal o número de iterações feitas. Gráfico esse representado nas figuras de 1A a 5B:

Figura 1A - Evolução da FO uchoa200 – Kmeans

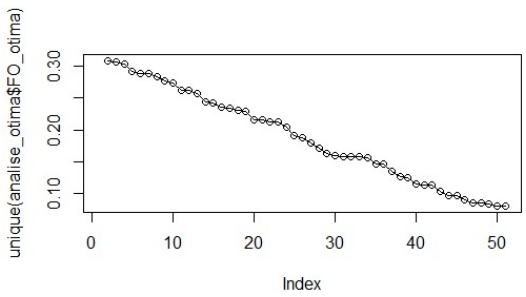


Figura 1B - Evolução da FO uchoa200 – SOM

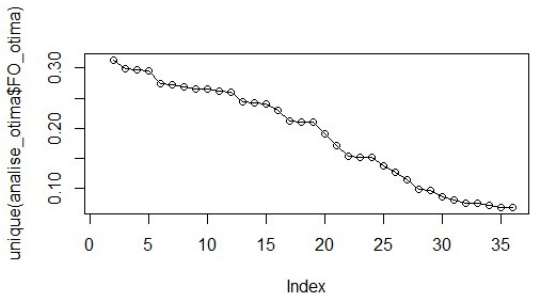


Figura 2A - Evolução da FO uchoa400 – Kmeans

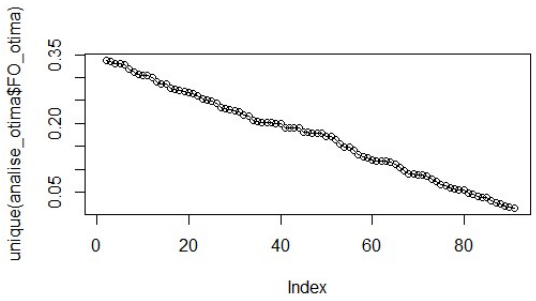


Figura 2B - Evolução da FO uchoa400 – SOM

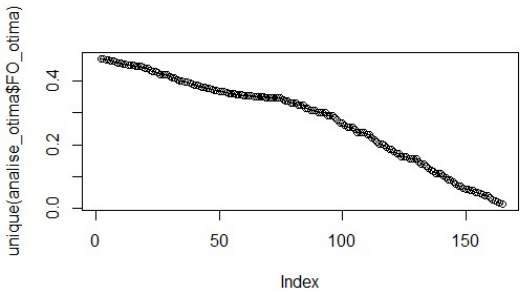


Figura 3A - Evolução da FO uchoa600 – Kmeans

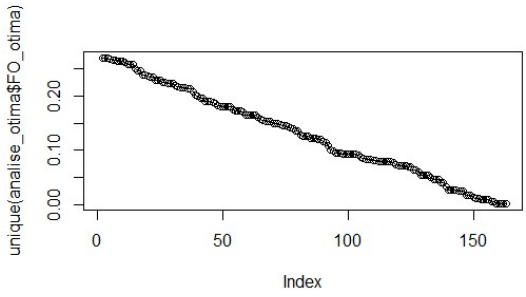


Figura 3B - Evolução da FO uchoa600 – SOM

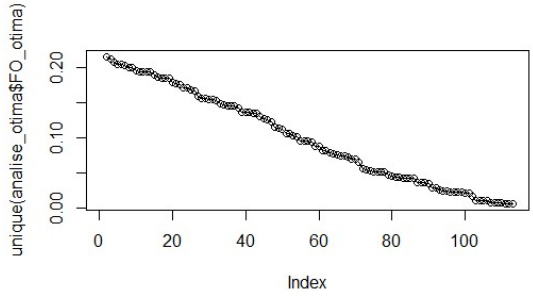


Figura 4A - Evolução da FO uchoa800 – Kmeans

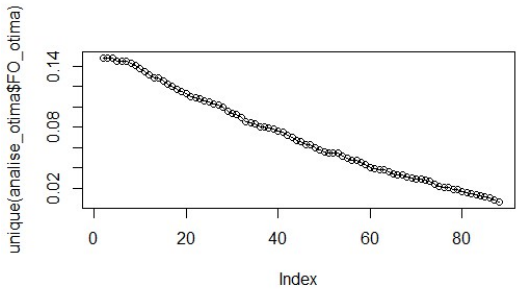


Figura 4B - Evolução da FO uchoa800 – SOM

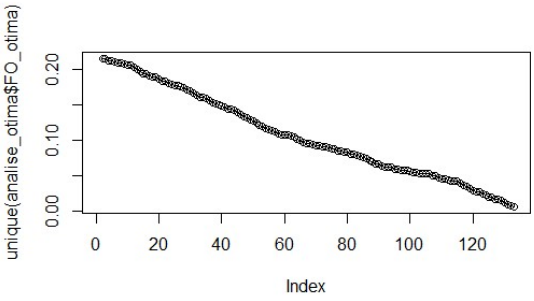


Figura 5A - Evolução da FO uchoa1000 – Kmeans

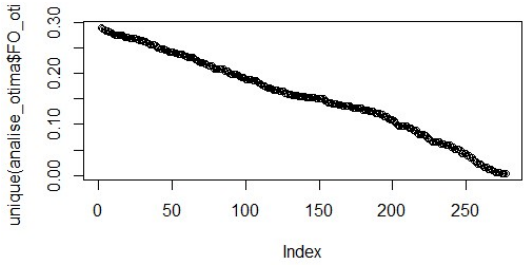
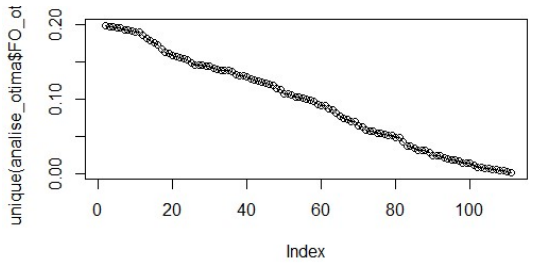


Figura 5B - Evolução da FO uchoa1000 – SOM



Um dos principais requisitos de um estudo que envolve heurísticas e metaheuristicas capazes de agrupar exemplares otimizando suas combinações é o seu resultado manter as características de cluster, ou seja, dados agrupados conforme suas similaridades. Como nesse projeto as similaridades são distribuídas em um plano bidimensional de coordenadas X e Y, os grupos devem envolver exemplares por proximidade e distância.

Os resultados obtidos em cada combinação neste estudo obtiveram êxito em cumprir esses requisitos. É possível observar que as clusterizações foram concretas e não houve desordem na distribuição dos exemplares em grupos conforme as figuras de 6A a 10B:

Figura 6A - Clusters Base uchoa200 – Kmeans

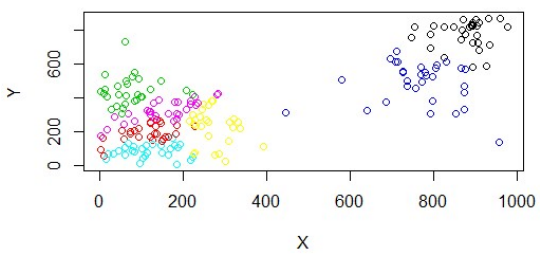


Figura 6B - Clusters Base uchoa200 – SOM

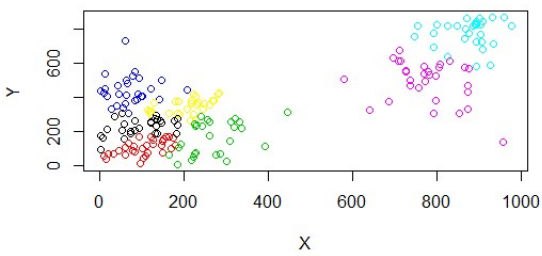


Figura 7A - Clusters Base uchoa400 – Kmeans

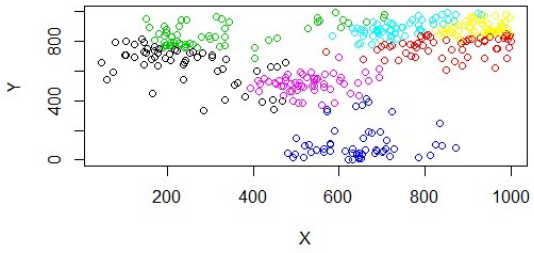


Figura 7B - Clusters Base uchoa400 – SOM

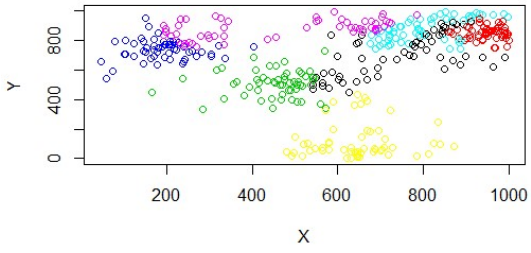


Figura 8A - Clusters Base uchoa600 – Kmeans

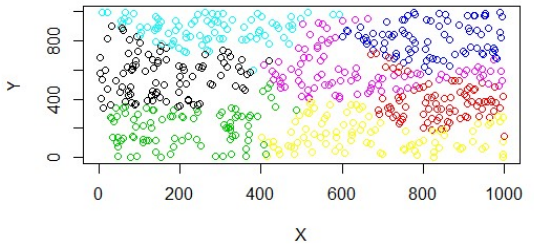


Figura 8B - Clusters Base uchoa600 – SOM

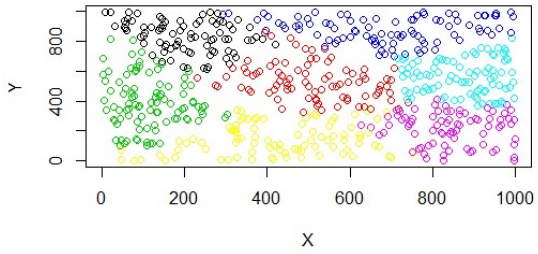


Figura 9A - Clusters Base uchoa800 – Kmeans

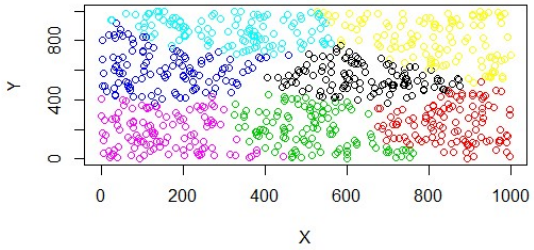


Figura 9B - Clusters Base uchoa800 – SOM

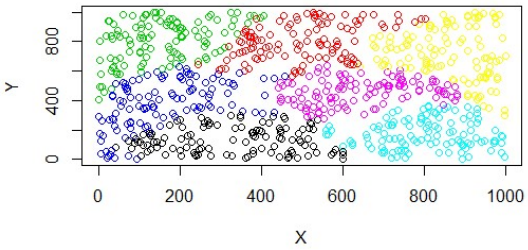


Figura 10A - Clusters Base uchoa1000 – Kmeans

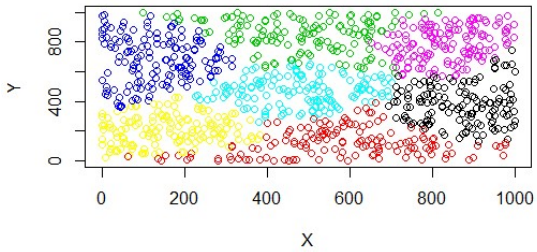
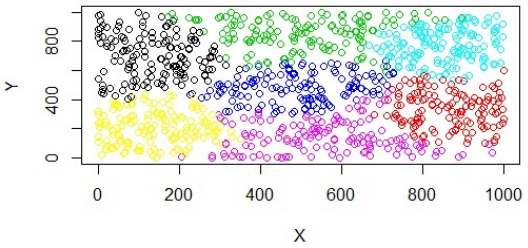


Figura 10B - Clusters Base uchoa1000 – SOM



5. CONSIDERAÇÕES FINAIS

Levando em consideração os resultados obtidos é possível concluir que os algoritmos de clusterização combinados com uma metaheurística de otimização combinatória é eficaz em termos de homogeneidade de distribuição de uma variável, que no caso da pesquisa faz alusão ao peso de cargas presentes na área de logística.

O desenvolvimento da pesquisa acarretou a criação de um novo método de

clusterização. Levando-se em consideração um ou mais fatores limitantes, as funções criadas são capazes de definir o rumo do algoritmo em tempo de execução, podendo gerar resultados adaptados aos parâmetros desejados.

Não há diferenciação direta entre os métodos de clusterização KMeans e Self Organizing Maps quando utilizada uma metaheuristica de otimização combinatória (*Simulated Annealing*). Porém vale ressaltar que o método KMeans sorteia aleatoriamente exemplares centroides para iniciar o processo de agrupamento e o SOM agrupa diretamente de acordo com suas características iniciais.

É possível gerar análises exploratórias com a base de dado resultante das funções desenvolvidas, podendo trazer informações dos clusters como número de exemplares, peso máximo e possíveis rotas para percorrer os exemplares.

As funções desenvolvidas se mostraram eficazes e prontas para serem aplicadas no cotidiano. Poderiam ser utilizadas para definir grupos de destinos de entregas levando-se em consideração a carga máxima de caminhões, segmentar cidades e ou bairros tendo em vista um orçamento total e uma quantidade mínima de exemplares e até classificar franquias de uma determinada empresa limitando cada grupo a soma de patrimônio de cada exemplar.

É recomendado como continuação dessa pesquisa a testagem e a análise de comportamento dos algoritmos desenvolvidos com bases contendo uma quantidade maior de registros e exemplares, assim é possível dizer se os métodos construídos são eficientes em situações globais ou de massa volumosa de dados.

Também serve de continuação a integração dos resultados obtidos com ferramentas de análises exploratória dos dados. Se feita de forma automática e fluida, o processo de clusterização torna-se mais completo, trazendo além do resultado desejado, que é a formação de grupos com características semelhantes, informações importantes e descritivas sobre os clusters formados. Estas informações podem ser exibidas em relatórios, dashboards e apresentações.

A utilização de outras técnicas de otimização combinatória também pode ser feita. A comparação entre elas pode trazer informações importantes como o surgimento de novas funções combinando os métodos clássicos com a nova técnica. Isso, se a técnica for mais eficaz, resulta numa possível melhor solução para o conjunto de dados.

REFERÊNCIAS

Alshamrani, R.; Alshehri, F.; Kurdi, H. (2020). “A Preprocessing Technique for Fast Convex Hull Computation”. The 11th International Conference on Ambient Systems, Networks

and Technologies (ANT). 6 a 9 de Abril, 2020. Varsóvia, Polônia. Procedia Computer Science 170, p. 317–324.

Brown, K. Q. (1979). "Voronoi diagrams from convex hulls". Information Processing Letters. Vol 9, No. 5

Colorni,A., Dorigo, M., Maffioli, F., Maniezzo, V., Righini, G., Trubian, M. (1997). “Heuristics from Nature for Hard Combinatorial Optimization Problems“; International Transactions in Operational Research, No.3.1 ,p. 1 - 38.

Crainic, T.G. e Laporte, G. (1997). “Planning Models for Freight Transportation“. European Journal of Operational Research 97, p. 409-438.

D’Amico, S. J.; Shoou-Jiun, W.; Batta, R.; Rump, C. M. (2002). “A simulated annealing approach to police district design”. Computers and Operations Research. Vol 29, pp. 667-684.

Daskin, M. S. (1995). “Network and discrete location: models, algorithms, and applications“. John Wiley & Sons Inc. N.Y.

Francis, R.L.; Rushton, G.; Rayco, M.B. (1999). "A synthesis of aggregation methods for multi-facility location problems: strategies for containing error," Geographical Analysis. 67-87,31.

Furuta, T.; Suzuki A.; Inakawa, K. (2005). The kth nearest network Voronoi diagram and its application to districting problem of ambulance systems. Technical Report 0501. Nanzan University.

Galvão, L.C.; Novaes, A. G.; Cursi, J. E. S.; Souza, J. C. (2006). A multiplicatively-weighted Voronoi diagram approach to logistics districting. Computers and Operations Research. Volume 33 , Issue 1 . pp: 93 – 114.

Glover, F. (1986). Future paths for integer methods for integer programming and links to artificial intelligence. Computers and Operations Research. n. 13, p. 533-549.

_____ (1990). Tabu search: a tutorial. Interfaces. July-August, 1990, pp. 74-94.

_____ (1993). A user’s guide to tabu search. Annals of Operations Research, 41, pp. 3-28.

Jayaram, M. A.; Fleyeh, H. (2016). “Convex Hulls in Image Processing: A Scoping Review”. American Journal of Intelligent Systems 2016, 6(2): p. 48-58.

Johnson, D.S. Aragon, C.R. McGeoch, L.A. Catherine, S. (1989 e 1991). “Optimization by simulated annealing: an experimental evaluation; part I e II“. Operations Research, Nov/Dec; 37, 6, p.865-892 e May/Jun; 39, 3, p.378-406.

Laporte, G.; Gendreau, M.; Potvin, J. Y. (1999). “Classical and modern heuristics for the vehicle routing problem“.Les Cahiers du Gerad G–99–21.

Matteucci, M. (2006). Tutorial in clustering. Politecnico di Milano. Department of Eletronics and Information. Milano, Italy. Disponível em 15/05/2006 no endereço: http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/

Novaes, A.G.; Graciolli, O. D. (1999). “Designing multi-vehicle delivery tours in a grid-cell format“. European Journal of Operational Research. Vol. 119, p. 613-614.

Novaes, A.G.; Cursi, J. E. S.; Graciolli, O. D. (2000). “A continuous approach to the designing of physical distribution systems“. Computers and Operations Research. Vol. 27, pp. 877-893.

Ochi, L. S., Dias, C. R., Soares, S. S. F. (2005). Clusterização em Mineração de Dados. Documento Técnico do Programa de Pós-Graduação em Computação. Instituto de Computação da Universidade Federal Fluminense. Niterói, RJ.

Pirlot, M. (1996). “General local search methods“. European Journal of Operational Research. Vol. 92, p. 493-511.

Silva, L. A.; Peres, S. M.; Boscarioli, C. (2016). “Introdução à Mineração de Dados: Com Aplicações em R”. 1ª Ed. Elsevier Editora Ltda. R.J. 277 p.

Vallim Fo , A. R. A. (2004) Localização de centros de distribuição de carga: contribuições à modelagem matemática. Tese de Doutorado, Escola Politécnica da Universidade de São Paulo. São Paulo, 286p..

Vallim Fo, A. R. A.; Gualda, N. D. F (2002). “A quantitative procedure for selecting best candidates for distribution center locations”. IMRL 2002 - Fourth International Meeting for Research in Logistics, 2002, Lisboa. v.3. p.887-896.

_____ (2004). “Heuristics, metaheuristics and optimization models for the distribution centers location problem”. XII Panamerican Conference on Traffic and Transportation Engineering, Albany, NY.

Zhou, G.; Hokey, M., Gen M. (2002). “The balanced allocation of customers to multiple distribution centers in the supply chain network: a genetic algorithm approach”. Computers and Industrial Engineering. V. 43, pp. 251-261.

Contatos: vgenerato11@gmail.com e arnaldo.aguiar@mackenzie.br