ISSN 2358-6036
v.9 2021
REVINT
REVISTA INTERDISCIPLINAR DE ENSINO, PESQUISA E EXTENSÃO
UNICRUZ

OTIMIZAÇÃO APLICADA A TÉCNICA DE CLUSTERIZAÇÃO

Optimization applied to the Clustering technique

Patricia Mariotto Mozzaquatro Chicon[1]

Alex Vinícios Telocken[2]

Resumo: Nos tempos atuais vivencia-se um crescente acúmulo de informações nas bases de dados das empresas. Para garantir sua permanência no mercado e competitividade, as mesmas buscam constantemente conhecimentos para fundamentar a tomada de decisão. Como muitas vezes esse conhecimento está oculto em uma grande base de dados, torna-se necessário a utilização de mecanismos computacionais que permitam a interação com estes dados de forma inteligente e rápida. Nesse sentido, a Mineração de Dados (MD) é uma alternativa que visa extrair conhecimento de um grande volume de dados, descobrindo novas correlações, padrões e tendências entre as informações de uma empresa. Este artigo tem por objetivo aplicar a técnica de clusterização, integrante da MD a fim de agrupar dados similares gerando informações relevantes. Foi aplicado o conceito de otimização, sendo utilizado o método Elbow a fim de determinar o número ótimo de cluster (K).

Palavras-chave: Otimizaçao. Clusterização. K-means. Elbow.

Abstract: Nowadays, there is a growing accumulation of information in companies' databases. To ensure their permanence in the market and competitiveness, they are constantly seeking knowledge to support decision making. As this knowledge is often hidden in a large database, it is necessary to use computational mechanisms that allow the interaction with this data in an intelligent and fast way. In this sense, Data Mining (DM) is an alternative that aims to extract knowledge from a large volume of data, discovering new correlations, patterns and trends between a company's information. This article aims to apply the clustering technique, part of MD, in order to group similar data generating relevant information. The concept of optimization was applied, using the Elbow method in order to determine the optimal cluster number (K).

Keywords: Optimization. Clustering. K-means. Elbow.

[1] Docente do Curso de Ciência da Computação. Universidade de Cruz Alta - Unicruz, Cruz Alta, Brasil. E-mail: patriciamozzaquatro@gmail.com

[2] Docente do Curso de Ciência da Computação. Universidade de Cruz Alta - Unicruz, Cruz Alta, Brasil. E-mail: telockenalex@unicruz.edu.br

1 INTRODUÇÃO

O volume de dados produzidos, armazenados ou transmitidos no mundo tem crescido exponencialmente nos últimos anos. Com a popularização do armazenamento na nuvem e a *Internet of Things* (IOT) o volume de dados para 2021 é previsto alcançar a marca de 40 zettabytes e o número de dispositivos conectados, gerando dados, chegará a 200 bilhões (ZWOLENSKI; WEATHERILL, 2014).

O surgimento dessa enorme quantidade de dados constitui o que se chama atualmente de Big Data, que de acordo com Tansley *et al.* (2009) é considerado o quarto paradigma da ciência. Big Data é certamente uma das expressões mais populares, impactante e mercadológico surgido nos últimos anos para referenciar uma nova área de pesquisa, cujos desafios envolvem a capacidade de trabalhar com a grande quantidade de dados gerados em todo mundo para ser acessado a uma maior velocidade (MACHADO, 2013). Na sociedade da informação atual, é fundamental saber tratar e analisar os dados em tempo hábil para que não se tornem inúteis.

Neste sentido, torna-se necessário a utilização de mecanismos que permitam entender e tirar um proveito maior dessas volumosas quantidades de dados. Frente a isso, uma alternativa são as ferramentas e técnicas de MD, que estão sendo cada vez mais empregadas em organizações e pesquisadas em ambiente acadêmico, pois oferecem de forma rápida, automatizada ou semi-automatizada (RAVAL, 2012). Uma alternativa para a geração de informações e produção do conhecimento, encontrando relacionamentos, padrões e tendências sobre os dados de uma forma que sejam úteis e possam auxiliar as tomadas de decisões dos mais diversos setores (HAND, 2007).

A pesquisa aqui apresentada tem por objetivo testar a técnica de clusterização, integrante da MD a fim de agrupar dados similares gerando informações relevantes. A fim de determinar o número ótimo de cluster (K) foi aplicado o método de otimização Elbow.

1.1 Mineração de dados

Na literatura é possível encontrar diversas definições para a MD, mas um dos principais conceitos, aceito por muitos pesquisadores, foi elaborado por Fayyad et.al, (1996) como: “o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis”. Esses autores referem-se, ainda, a Descoberta do Conhecimento em Bases de Dados (DCBD) como um processo global de descoberta de conhecimento que envolve seleção, pré-processamento dos dados e

transformação dos mesmos, também mineração de dados, interpretação dos resultados e a transformação do conhecimento. A MD é uma das etapas deste processo onde são aplicados algoritmos específicos para extração de padrões a partir dos dados ou até mesmo revelar o comportamento de um banco de dados.

Para Han e Kamber (2006), a MD também pode ser descrita como uma área de pesquisa multidisciplinar que engloba diversas outras áreas como: Inteligência Artificial; Aprendizado de Máquina; Redes Neurais; Estatística; dentre outras. Além disso, de acordo com Goldschmidt et.al, (2015) duas novas áreas de MD fortemente relacionadas com o Big Data são *Parallel Data Mining* (PDM) e *Distributed Data Mining* (DDM). As principais diferenças envolvem escala, custos de comunicação e distribuição de dados. "DDM pode dispor de um número muito superior de processadores que PDM. O custo de movimentação de dados em DDM é superior quando comparado ao custo de movimentação de dados em PDM". Os autores destacam que a PDM é a escolha adequada para bases de dados centralizadas, enquanto DDM é mais indicada em casos onde há múltiplas bases de dados distribuídas.

Conforme os autores Han e Kamber (2006), dentre as tarefas mais comuns de mineração de dados é possível citar: Classificação, Agrupamento, Associação e Regressão.

Classificação: utiliza aprendizado supervisionado. Nesta tarefa os atributos do conjunto de dados são divididos em dois grupos, ou seja, um dos grupos apresenta somente um atributo (atributo alvo) para o qual se deve fazer a predição de um valor. Quando se aplica a técnica de classificar, o atributo alvo é categórico. O outro grupo apresenta os atributos a serem utilizados na predição (GOLDSCHMIDT *et al.*, 2015).

Clusterização: também é chamada de agrupamento ou análise de agrupamento, utiliza aprendizado não supervisionado. Objetiva-se separar os registros de um conjunto de dados em subconjuntos ou grupos (clusters) de tal forma que elementos em um cluster compartilhem um conjunto de propriedades comuns que os distinguem dos elementos de outros clusters) (GOLDSCHMIDT *et al.*, 2015).

Associação: também denominada de regras de associação. Genericamente uma regra de associação é representada pela notação X -> Y (X implica em Y), onde X e Y são conjuntos de itens distintos (DIAS, 2001). Intuitivamente essa tarefa objetiva encontrar conjuntos de itens que ocorram simultaneamente e de forma frequente em um conjunto de dados (GOLDSCHMIDT *et al.*, 2015). A tarefa de associação se enquadra no modelo descritivo, buscando revelar ocorrências frequentes, tendências e/ou padrões nos dados.

Regressão: compreende, fundamentalmente, a busca por funções, lineares ou não, que mapeiem os dados de uma base de dados em valores reais. A regressão é uma tarefa semelhante à classificação, entretanto, sua aplicação, restringe-se apenas a dados do tipo numérico (GOLDSCHMIDT *et al.*, 2015).

Segundo Dias (2001), a escolha das técnicas de MD dependerá da tarefa específica a ser executada e dos dados disponíveis para análise, devendo-se levar em conta a natureza dos dados disponíveis em termos de conteúdo, os tipos de campos de dados e a estrutura das relações entre os registros." A pesquisa aqui apresentada irá abordar a técnica de clusterização.

1.1.1 Técnica de Clusterização

A técnica de clusterização também é chamada de agrupamento ou análise de agrupamento, utiliza aprendizado não supervisionado. Objetiva-se separar os registros de um conjunto de dados em subconjuntos ou grupos (clusters) de tal forma que elementos em um cluster compartilhem um conjunto de propriedades comuns que os distinguem dos elementos de outros clusters (GOLDSCHMIDT *et al.*, 2015).

Neste sentido a tarefa de agrupar pode ser interpretada como um problema de otimização, ou seja, objetiva-se maximizar a similaridade intracluster e minimizar a similaridade intercluster. Existem na literatura diversos algoritmos de clusterização, neste artigo será abordado o algoritmo k-Means. O algoritmo de clusterizacao, K-Means (também chamado de K-Médias) fornece um agrupamento de informações de acordo com os próprios dados. O algoritmo K-Means é popular devido a sua facilidade de implementação e sua ordem de complexidade O (n), onde n é o número de padrões (FONTANA *et al*., 2009).

No agrupamento, os objetos considerados como entrada não possuem rótulos associados. A clusterização analisa grupos por meio de um conjunto de objetos em grupos de acordo com alguma medida de similaridade.

1.1.2 Método K-Means

K-Means utiliza o conceito de centróides como protótipos representativos dos grupos, onde o centróide representa o centro de um grupo, sendo calculado pela média de todos os objetos do grupo. O processo iterativo termina quando os centroides dos grupos param de se modificar, ou após um número preestabelecido de iterações ter sido realizadas.

O algoritmo de K-Means pode ser visto através dos passos definidos por Fontana *et al.* (2009):

1. Atribuem-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo;

2. Atribui-se cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto;

3. Recalcula-se o valor do centróide (protótipo) de cada grupo, como sendo a média dos objetos atuais do grupo; repete-se os passos 2 e 3 até que os grupos se estabilizem.

Existem vários algoritmos k-Means disponíveis. O algoritmo padrão é o Algoritmo de Pérez-Suárez *et al.* (2014), que define a variação total dentro do cluster como a soma das distâncias quadradas, utiliza a distância euclidiana entre os itens e os correspondentes centroides, descrito por $W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$ onde xi serve para projetar um ponto de dados pertencente ao cluster Ck; uk é o valor médio dos pontos atribuídos ao cluster Ck; cada observação (xi) é atribuída a um determinado cluster de modo que a soma dos quadrados e a distância da observação aos centros de cluster designados uk é um mínimo. A variação total dentro do cluster, é definida por $tot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$

A soma total dentro do cluster do quadrado mede a compactação (ou seja, a eficácia) do agrupamento, ou seja, deve ser a menor possível. O algoritmo K-means consiste em quatro etapas:

Inicialização: o analista determina o número k de grupos desejados e, em seguida, o algoritmo define um centroide para cada um. Algumas implementações do algoritmo têm vários tipos de inicialização, entre as quais foram avaliadas para possível implementação, conforme (Pérez-Suárez *et al.*, 2014):

Random: os centroides k iniciais são determinados aleatoriamente;

K-Means++: o procedimento após este tipo de inicialização para corrigir os Centroides é: 1. definir o primeiro centroide $\boldsymbol{c1}$. Este é escolhido aleatoriamente em um conjunto de dados. 2. determina-se o novo centroide $\boldsymbol{cx}$, escolhendo o elemento de conjunto de dados mais provável. 3. iterar na etapa 2 até obter os centroides K de cada grupo.

Canopy: O primeiro centroide é corrigido aleatoriamente e, em seguida, executa uma primeira partição usando uma métrica mais simples, em nível computacional, em que os grupos são gerados em sobreposição.

1.1.2.1 Determinação do número ótimo de cluster

Determinar o número ideal de clusters em um conjunto de dados é fundamental a fim de resolver o problema no particionamento de clustering. Ao se aplicar a clusterização se requer que o usuário especifique o número de clusters (k) a serem gerados. Conforme estudos já realizados por Kassambara (2017), não há uma resposta definitiva para essa questão. O número ideal de clusters é de alguma forma subjetiva e depende do método utilizado para medir semelhanças e os parâmetros usados para o particionamento.

Neste contexto, a literatura aborda o método Elbow, ou seja, examina o erro quadrado total como uma função do número de clusters.

O método Elbow trata-se de uma técnica para encontrar o valor ideal do parâmetro k. Basicamente o que o método faz é testar a variância dos dados em relação ao número de clusters (KASSAMBARA, 2017). É considerado um valor ideal de k quando o aumento no número de clusters não representa um valor significativo de ganho.

O método aplicado na Figura 1, segue os procedimentos:

1.Calcular o algoritmo de clusterização k-Means para diferentes valores de k;

2. Para cada k, calcular a soma total dentro do cluster do erro quadrado;

3. Plotar a curva do erro quadrado de acordo com o número de clusters k;

4. A localização de uma curva é considerada como indicador do número apropriado de clusters, ou seja, quando as distâncias dos erros quadráticos praticamente se estabilizam.

1.2 Otimização

A otimização matemática é uma área da ciência computacional que busca responder a pergunta “O que é melhor?” para problemas em que a qualidade de uma resposta pode ser medida por um número. Estes problemas aparecem em praticamente todas as áreas do conhecimento: negócios, ciências físicas, químicas e biológicas, engenharia, arquitetura, economia e administração (HILLIER *et al.*, 2013).

Os problemas de otimização são problemas de maximização ou minimização de função de uma ou mais variáveis em um determinado domínio, sendo que, geralmente, existe um conjunto de restrições nas variáveis (BASTOS-FILHO *et al.*, 2009).

Tratando-se da otimização, é necessário definir alguns conceitos, conforme o autor Bastos-Filho *et al.* (2009):

- Variáveis de projeto: São aquelas que se alteram durante o processo de otimização, podendo ser contínuas (reais), inteiras ou discretas;
- Restrições: São funções de igualdade ou desigualdade sobre as variáveis de projeto que descrevem situações de projeto consideradas não desejáveis;
- Espaço de busca: É o conjunto, espaço ou região que compreende as soluções possíveis ou viáveis sobre as variáveis do projeto do problema a ser otimizado, sendo delimitado pelas funções de restrição;
- Função Objetivo: É a função de uma ou mais variáveis de projeto que se quer otimizar, minimizando-a ou maximizando-a;
- Ponto Ótimo: É o ponto formado pelas variáveis de projeto que extremizam a função objetivo e satisfazem as restrições;
- Valor Ótimo: É o valor da função objetivo no ponto ótimo.

Os métodos de otimização podem ser divididos em dois grupos: Programação Linear e Programação Não Linear. Programação Linear: tem como objetivo encontrar a solução ótima em problemas nos quais a função objetivo e todas as restrições são representadas por funções lineares das variáveis do projeto (GOLDBARG *et al.*, 2017). Programação Não Linear: trata dos problemas em que a função objetivo ou algumas das restrições são funções não lineares das variáveis envolvidas. Os métodos de Programação Não Linear são classificados em dois subgrupos: Métodos Determinísticos e Métodos Não Determinísticos (GOLDBARG *et al.*, 2017).

Métodos Determinísticos: os métodos de otimização baseados nos algoritmos determinísticos geram uma sequência determinística de possíveis soluções requerendo, na maioria das vezes, o uso de pelo menos a primeira derivada da função objetivo em relação às variáveis de projeto (HILLIER *et al.*, 2013). Um método de otimização é chamado de Determinístico se for possível prever todos os seus passos conhecendo seu ponto de partida.

Métodos Não Determinísticos: procuram imitar fenômenos ou processo encontrados na natureza e, por este motivo, pertencem a uma área denominada computação natural.

2 METODOLOGIA

Esta pesquisa tem como objetivo aplicar a técnica de clusterização a fim de agrupar dados similares utilizado o método Elbow. O método irá determinar o número ótimo de cluster (K).

Quanto a natureza, a pesquisa classifica-se como aplicada. Segundo Carmo *et al.* (2015), a pesquisa aplicada é definida pela necessidade do pesquisador solucionar um problema existente ou não, e com isso obter resultados.

Quanto aos procedimentos, classifica-se como pesquisa experimental. Para Gil (2002) a pesquisa experimental consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto. A pesquisa foi realizada nas seguintes etapas:

Etapa 1: Estudo Teórico sobre: Mineração de Dados, Técnica de Clusterização, Método K-Means e software Waikato Environment for Knowledge Analysis (Weka), Otimização e Método Elbow.

Etapa 2: definição das variáveis de decisão:

-métodos de inicialização: k-means++, canopy e randow;

-número de cluster;

-restrição: <=500 iterações;

Função objetivo: encontrar o melhor número de cluster em função da variação dos erros quadrados;

Parâmetros: erro quadrado, cluster, iterações.

Etapa 3: Criação das bases de dados: a base de dados foi gerada utilizando o banco de dados Mysql. Foram geradas bases randômicas nos seguintes tamanhos: 1000, 10.000, 100.000, 200.000, 300.000, 400.000 e 500.000. Trabalhou-se com 7 atributos (idade, peso, altura, IMC, circunferência da cintura, pressão sistólica e pressão diastólica).

Após a geração das bases, exportou-se as mesmas na extensão ".csv".

Etapa 4: conversão da base "csv" para "arff". Realizou-se a conversão a fim de integrar a mesma ao software Weka.

O software Weka é uma coleção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente a uma série de dados ou ser chamados de seu próprio código Java. Weka contém ferramentas para o pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização. E também pode ser usada para desenvolver novos algoritmos de aprendizagem de máquina (BOUCKAERT *et al.,* 2010).

O software apresenta quatro opções de trabalho: Explorer (nela é possível aplicar as tarefas e métodos sobre a base de dados), Experimenter (consiste em aplicar um ou vários métodos de classificação sobre uma grande quantidade de dados, além de ter condições de

realizar comparações estatísticas) knowledgeFlow (considerada a interface que apresenta de forma mais explicita o funcionamento da ferramenta, tendo sua representação na forma gráfica) e Simple CLI (proporciona um local para inserir comandos (BOUCKAERT *et al.*, 2010).

Etapa 5: Carregar a base no software Weka.

O formato de arquivo ARFF é um arquivo que contém texto, sendo dividido por três partes: A primeira linha do arquivo deve ser igual a @relation. Na parte seguinte uma palavra-chave que identifique a relação ou tarefa sendo estudada. E após, aparece o Atributo que é um grupo de linhas onde cada uma delas é iniciada com @attribute seguido do nome do atributo, depois seu tipo. O tipo pode ser nominal; as alternativas são apresentadas em listas separadas por virgula e cercada por chaves ou, poderá ser numérico mas o nome deverá ser precedido pela palavra-chave real. Logo em seguida uma linha será contida por @data. Cada linha condiz a uma instancia a qual deve ter valores separados por vírgula assim correspondendo a mesma ordem dos atributos da seção atributos. A linha que contiver na parte inicial o sinal de porcentagem (%), será contada como comentário e não será processada.

Etapa 6: Aplicação da técnica e método de clusterização no software Weka. Ao escolher a opção Explorer uma nova tela é apresentada, com as técnicas e algoritmos de clusterização, seleção dos atributos e suas visualizações.

Com os dados já formatados, definiu-se a técnica de clusterização e o método k-Means a ser aplicado o agrupamento. Após definiu-se os parâmetros para a geração da informação. Como método de inicialização trabalhou-se com: K-means++, canopy e Randow. Como distância, optou-se pela distância euclidiana.

Para poder agrupar conjuntos de objetos é necessário medir a similaridade entre eles. Esta medida é obtida com um cálculo de distância entre os objetos. A pesquisa aqui apresentada utilizou a distância euclidiana $d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$

Definiu-se como número máximo de iterações 500. Trabalhou-se com valores de cluster (K) de 1 a 10

Etapa 7: geração da informação.

3 RESULTADOS E DISCUSSÃO

A seção aqui apresentada irá descrever os resultados gerados após a execução da técnica de clusterização no software Weka. Os resultados são descritos por meio de análise gráfica.

Primeiramente analisou-se o squared erros validando o melhor método de inicialização: k-means++, canopy e randow. Analisou-se com 2,4,6,8,10 cluster. A Tabela 1 ilustra os resultados das bases de tamanhos 10.000, 100.000, 200.000, 300.000, 400.000 e 500.000.

Tabela 1- Squared erros: 10.000,100.000,200.000, 300.000, 400.000 e 500.000.

		Squared Errors – 10.000	
cluster	Canopy	K-means++	Randow
2	4.225,8821844482100	4.225,8822959197700	4.225,8822959197700
4	3.406,3119693133400	3.385,9053648564000	3.406,2849861517700
6	2.956,4886166453400	2.944,9274394246800	2.953,8064103787900
8	2.646,3484972581400	2.658,7963200201100	2.659,3853187087800
10	2.398,8180229107500	2.411,0639740804600	2.403,7724768697300
		Squared Errors – 100.000	
cluster	Canopy	K-means++	Randow
2	48.326,4593792955000	48.326,4593468136000	48.326,4594297028000
4	38.740,6999409511000	39.379,2009251785000	38.740,7159822267000
6	33.813,8244136950000	33.784,0051667899000	33.928,9565615385000
8	30.550,5841655280000	30.332,8170376541000	30.192,1357292781000
10	27.585,1401392308000	27.604,8986734179000	27.520,2708022989000
		Squared Errors – 200.000	
cluster	Canopy	K-means++	Randow
2	85.501,9030081259000	85.501,9030081259000	85.501,9030081259000
4	68.609,8043094301000	68.609,8062781432000	68.609,8042922433000
6	60.480,4405224996000	59.961,8500556227000	59.957,5199636925000
8	53.779,8403659526000	53.723,9184248505000	53.560,2964545461000
10	48.793,0717654293000	48.795,1530574621000	48.803,5001388327000
		Squared Errors – 300.000	
cluster	Canopy	K-means++	Randow
2	128.290,0960865360000	128.290,0957742500000	130.939,7827415900000
4	104.616,7429099860000	104.551,6702206260000	102.930,5997087950000
6	90.562,1226345761000	90.433,9494169276000	90.120,7160336178000
8	80.941,2728263834000	81.011,9249478311000	80.593,8815204284000
10	73.250,3564138796000	73.373,3639618513000	73.170,8386625294000
		Squared Errors – 400.000	
cluster	Canopy	K-means++	Randow
2	171.169,7087167510000	171.169,7087167510000	174.712,7947246370000
4	137.304,3061963610000	137.304,3062778710000	137.304,3456385980000
6	119.795,6025485610000	120.271,3875650940000	119.771,6806312890000
8	107.151,6989776560000	107.151,6746584660000	107.168,1913690460000
10	não executa	97.628,7336404425000	97.630,3254046229000
		Squared Errors – 500.000	
cluster	Canopy	K-means++	Randow
2	214.186,2807608160000	214.186,2807608160000	218.550,9802213960000
4	171.765,6028174870000	171.765,5991165090000	171.765,5991165090000
6	149.854,8866651980000	149.854,8768364920000	149.854,8781780720000
8	134.523,8139963010000	134.518,1412874100000	134.554,6026188970000
10	não executa	122.177,6940517520000	122.108,9992333050000

Fonte: Autores (2021).

Após análise realizada constatou-se que para bases de tamanho menores que 10.000 o método de inicialização k-means++ obteve melhor desempenho no agrupamento gerado, isto é, obteve-se um menor valor para squared erros. Em contrapartida, para bases maiores que 10.000 o método Randow obteve melhor desempenho (100.000, 200.000, 300.000, 400.000 e 500.000).

A escolha do método de inicialização Randow, para a etapa seguinte, justifica-se devido aos testes realizados, pois o objetivo é trabalhar com bases Big Data (grande quantidade de dados).

A etapa seguinte refere-se a escolha do número de cluster ótimo. Com o método de inicialização já definido (método Randow), realizou-se mais testes optando por valores de 1 a 10 cluster utilizando o software Weka para a execução do algoritmo K-Means. A Figura 1 ilustra a variância relacionada aos squared erros gerados em relação ao número de cluster. Para determinar o K ótimo utilizou-se o método Elbow.

Conforme já descrito, o método Elbow define o número ótimo de K quando as distâncias dos erros quadráticos praticamente se estabilizam. Sendo assim, chegou-se as seguintes conclusões:

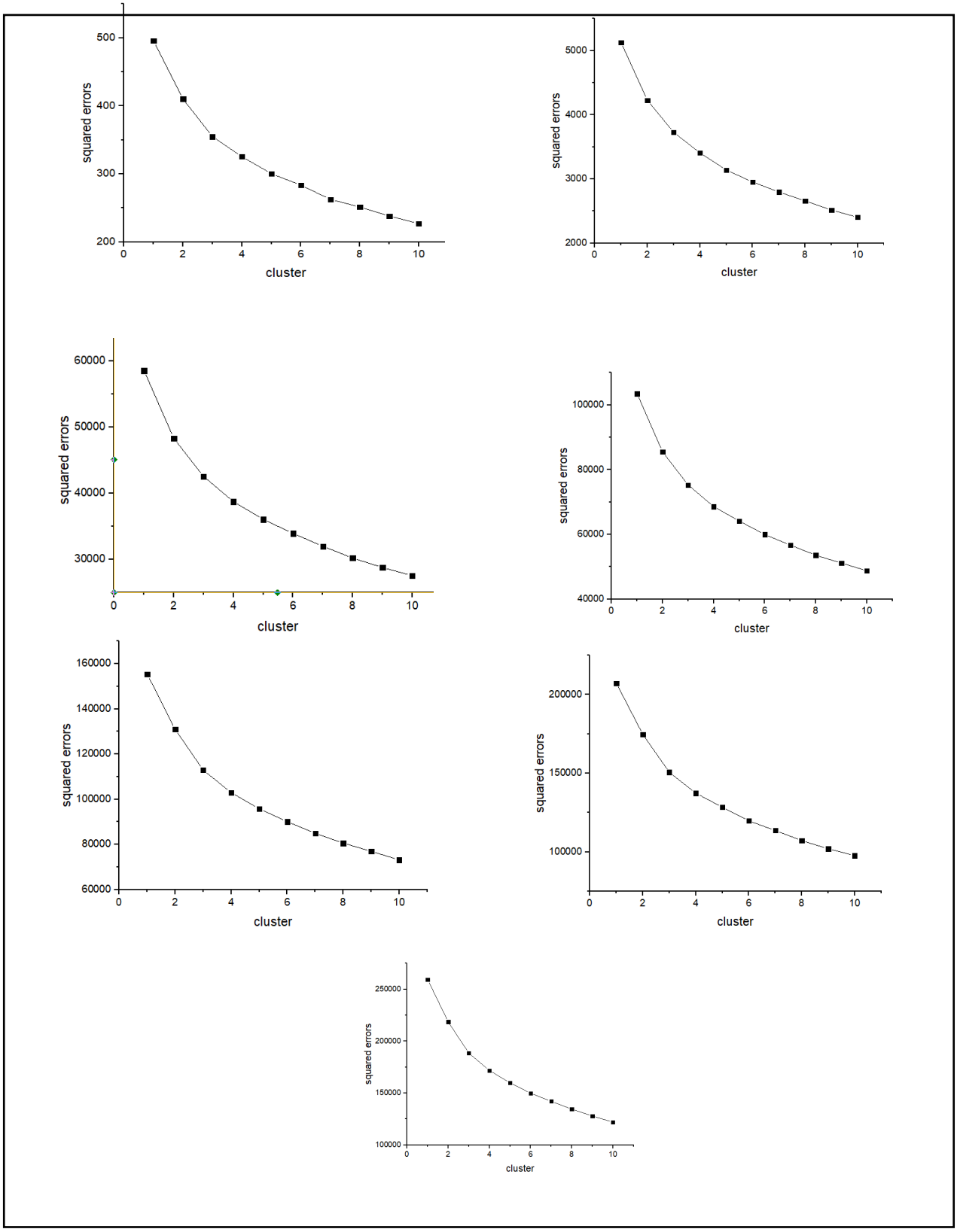
-em bases de tamanho entre 10.000 e 90.000 o número ótimo de K seria de até 5 cluster;

-em bases de tamanho maiores que 90.000 o número ótimo de k seria de até 6 cluster.

Observa-se que a partir do número de seis clusters as distâncias dos erros quadráticos praticamente se estabilizam, sendo assim não irá ocorrer aumento na eficácia do agrupamento com k maiores que 6.

Com os resultados obtidos, após os testes realizados com bases de tamanhos 1000 a 500.000 pode-se constatar que quando se está trabalhando com valores numéricos e objetiva-se aplicar a técnica de clusterização, integrando o método K-Means a fim de agrupar dados similares, para valores maiores que 90.000 o número de cluster ótimo será de 6.

Figura 1 - Squared Errors 1000, 10000, 100000, 200000, 300000, 400000, 500000.



Fonte: Autores (2021).

4 CONSIDERAÇÕES FINAIS

A pesquisa aqui apresentada buscou aplicar a técnica de clusterização validando o método k-Means. Teve por objetivo encontrar o número ótimo de cluster (K). Para este propósito foi utilizado o método estatístico Elbow. A base de dados foi gerada utilizando o banco de dados Mysql. Os seguintes tamanhos foram validados: 1000, 10.000, 100.000, 200.000, 300.000, 400.000 e 500.000. A execução do algoritmo de clusterização k-Means aconteceu por meio do software Weka. Validou-se o melhor método de inicialização. O experimento foi testado com os métodos de incialização K-means++, canopy e Randow. Como distância, optou-se pela distância euclidiana. Analisou-se o squared erros validando o melhor método de inicialização. Constatou-se que para bases de tamanho menores que 10.000 o método de inicialização k-means++ obteve melhor desempenho. Em contrapartida, para bases maiores que 10.000 o método Randow obteve melhor desempenho. Assim, como o objetivo da pesquisa é trabalhar com bases Big Data, ou seja, com grandes bases de dados, optou-se trabalhar com o método de inicialização Randow. Após a análise do experimento, também se constatou que o número de cluster ótimo para bases de bases de tamanho entre 10.000 a 90.000 o número ótimo de K seria de até 5. Em bases de tamanho maiores que 90.000 o número ótimo de k seria de até 6. Com os testes realizados, concluiu-se que a partir do número de seis clusters as distâncias dos erros quadráticos praticamente se estabilizam.

A principal contribuição deste trabalho foi determinar o número ótimo de cluster (K) a ser definido automaticamente pelo método k-Means integrante da técnica de clusterização. Ao se definir os parâmetros ideais do método, gera-se o agrupamento com alto grau de precisão. Também identificou-se uma série de desafios para pesquisas futuras relacionadas a eficácia da técnica de clusterização. Na pesquisa aqui abordada trabalhou-se com dados estruturados no formato numérico. Pretende-se validar com dados semiestruturados e não estruturados (numérico e textual).

REFERÊNCIAS

BASTOS-FILHO, C. J. A.; CARVALHO, D. F.; CARACIOLO, M. P.; MIRANDA, P. B.; FIGUEIREDO, E. M. Multi-ring particle swarm optimization. In: **Evolutionary Computation**. IntechOpen, 2009.

BOUCKAERT, R. R.; FRANK, E.; HALL, M.; KIRKBY, R.; REUTEMANN, P.; SEEWALD, A.; SCUSE, D. **WEKA manual for version 3-7-3**. The university of WAIKATO, v. 327, 2010.

CARMO, Liege Moraes do; MACHADO, Rodrigo Sahagoff; COGAN, Samuel. Uma análise do processo de elaboração do trabalho de conclusão de curso a partir do processo de raciocínio da teoria das restrições. **ReCont (Registro Contábil)**, v. 6, n. 3, 2015.

DIAS, Maria Madalena. **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados**. 2001.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

FONTANA, André; NALDI, M. C. **Estudo e comparação de métodos para estimação de números de grupos em problemas de agrupamento de dados**. 2009.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

GOLDBARG, Elizabeth; GOLDBARG, Marco; LUNA, Henrique. **Otimização Combinatória e Metaheurísticas: Algoritmos e Aplicações**. Elsevier Brasil, 2017.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data Mining**. Elsevier Brasil, 2015.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques**. San Francisco. Morgan Kalfmann Publishers, Secound Edition, 2006

HAND, David J. Principles of data mining. **Drug safety**, v. 30, n. 7, p. 621-622, 2007.

HILLIER, Frederick S.; LIEBERMAN, Gerald J. **Introdução à pesquisa operacional**. McGraw Hill Brasil, 2013.

KASSAMBARA, Alboukadel. **Practical guide to cluster analysis in R**: Unsupervised machine learning. Sthda, 2017.

MACHADO, F. **Tecnologia e Projeto de Data Warehouse: uma visão multidimensional**. 6ª Ed. São Paulo: Érica, 2013

PÉREZ-SUÁREZ, A.; MEDINA-PAGOLA, J. E. **Algoritmos para el agrupamiento conceptual de objetos**. La Habana: sn, 2014.

RAVAL, Kalyani M. Data mining techniques. **International Journal of Advanced Research in Computer Science and Software Engineering**, v. 2, n. 10, 2012.

TANSLEY, Stewart; TOLLE, K. M. **The fourth paradigm: data-intensive scientific discovery**. Redmond, WA: Microsoft research, 2009.

ZWOLENSKI, Matt; WEATHERILL, Lee. The digital universe: Rich data and the increasing value of the internet of things. **Journal of Telecommunications and the Digital Economy**, v. 2, n. 3, p. 47.1-47.9, 2014.