Um Estudo Sobre Processos para Avaliação de Algoritmos de Agrupamento de Dados

Aline M. M. Kronbauer, Lisandra Manzoni Fontoura, Ana Trindade Winck

Grupo de Pesquisa em Sistemas Inteligentes - Programa de Pós-graduação em Informática – Universidade Federal de Santa Maria (UFSM)

Caixa Postal 15.064 – 97105-900– Santa Maria – RS – Brasil

{alyne.k, lisandramf, anawinck}@gmail.com

Abstract. *This article presents the idea that assessment data clustering algorithms is a task that can be addressed using valuation methods oriented goals. The approach presented is based on GQM methodology (Goal, Question, Metric) for evaluation of processes and software products. From a certain data clustering problem, it uses the GQM approach to structure goals, questions, measures and metrics allowing for better structuring of goals to be achieved which facilitates the evaluation of data clustering algorithms. A test application of the proposed approach is presented, it the k-means algorithm is used.*

Resumo. *O presente artigo apresenta a ideia de que avaliação de algoritmos de agrupamento de dados é uma tarefa que pode ser abordada utilizando métodos de avaliação orientados a objetivos. A abordagem apresentada baseia-se na metodologia GQM (Goal, Question, Metric) para avaliação de processos e produtos de software. A partir de um determinado problema de agrupamento de dados, utiliza-se a abordagem GQM para estrutura objetivos, perguntas, medidas e métricas permitindo uma melhor estruturação das metas a serem alcançadas o que facilita na avaliação de algoritmos de agrupamento de dados. Um teste de aplicação da abordagem proposta é apresentado, nele o algoritmo k-means é utilizado.*

1. Introdução

Diversas bases de dados têm aumentado seu tamanho, isto se deve, em grande parte, à automatização de alguns processos e ao advento da rede mundial de computadores (WEB). Dados biológicos ou obtidos através de sensores, dados advindos das redes sociais e registros clínicos são alguns dos processos que tem seu tamanho ampliado com o passar do tempo.

A análise de grupos é uma área da mineração de dados que busca semelhanças ou diferenças em um determinado conjunto de dados e baseado nestes preceitos, realiza o agrupamento ou não desses dados. O agrupamento de dados é uma técnica não supervisionada (na terminologia de Data Mining), trata-se de uma atividade cujos dados não possuem rótulos, cujo resultado não é apto para receber comparações.

Há então que procurar outros meios para aferir a qualidade de um agrupamento. Nesse sentido, entende-se que o conhecimento dos principais algoritmos de

agrupamento de dados é uma questão muito importante para o correto direcionamento de um projeto voltado para essa área. Este projeto propõe-se realizar um levantamento de informações acerca de algoritmos de agrupamento de dados, investigar esses algoritmos e processos que envolvem sua avaliação e buscar formas para que seja possível a avaliação de forma a contribuir para o aumento da qualidade dos grupos resultantes.

Na tentativa de buscar a elaboração de um processo avaliativo para algoritmos de agrupamento de dados, o método GQM de avaliação é estudado como uma possível abordagem avaliativa. Processos de avaliação destes algoritmos serão estudados e uma proposta baseada na abordagem Goal, Question, Metric será apresentada, bem como um teste de aplicação com o algoritmo K-means.

2. Agrupamento de Dados

Agrupamento de dados busca, dado um conjunto de objetos, coloca-los em grupos (clusters) baseado na similaridade entre objetos. Para isso, deve-se maximizar a similaridade (homogeneidade) entre os objetos de um mesmo grupo e minimizar a similaridade entre objetos de grupos distintos Goldschimidt (2005), Halkidi (2001) e Larose (2005). É uma técnica não supervisionada de dados pois a similaridade entre os atributos é uma característica intrínseca dos dados por não precisar de um arquivo de treinamento com classes pré-definidas.

O objetivo de uma técnica de agrupamento é encontrar uma estrutura de clusters (grupos) nos dados em que os objetos pertencentes a cada cluster compartilham alguma característica ou propriedade relevante para o domínio do problema em estudo Faceli et al. (2011). É uma técnica não supervisionada de dados pois a similaridade entre os atributos é uma característica intrínseca dos dados por não precisar de um arquivo de treinamento com classes pré-definidas.

Em se tratando de grandes volumes de dados, muitas vezes, é impossível exigir uma rotulação, pois esta medida resulta em muitos gastos, tempo de aplicação e esforço. Para driblar este problema, adota-se a medida de trabalhar com dados não rotulados a fim de encontrar padrões existentes nos mesmos.

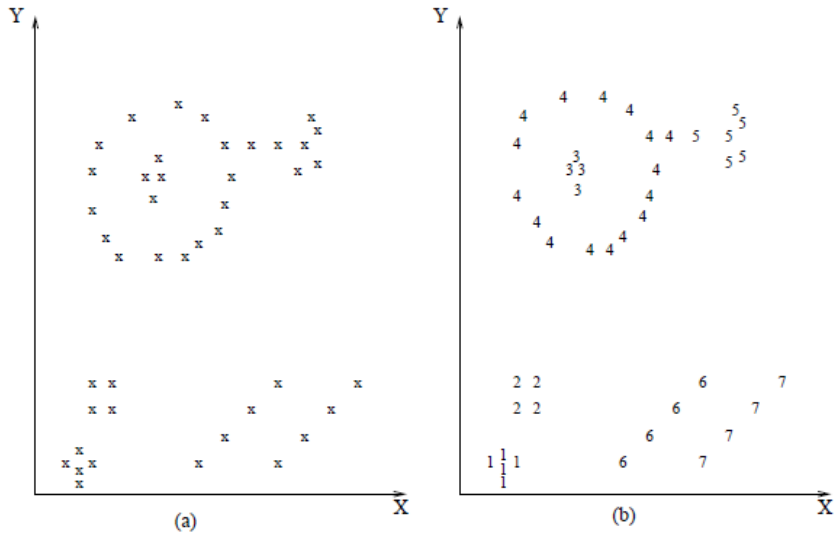

Figura 1: Agrupamento de dados Jain et al. (1999).

A Figura 1 traz um exemplo de agrupamento de dados. Nela, dois estágios são representados: o conjunto de dados de entrada (a) e o agrupamento resultante (b). O estágio (b) mostra os dados já agrupados depois de processados por determinado algoritmo. O processo de agrupamento resultou em 7 grupos, onde cada grupo é representado por um número, os dados que fazem parte do mesmo grupo possuem o mesmo número.

Agrupamento de dados é considerada uma das tarefas mais úteis no processo de mineração de dados, é utilizada para descobrir grupos e identificar distribuições interessantes e padrões nos dados subjacentes. O tema pode ser encontrado com diferentes nomes e em diferentes contextos, tais como aprendizagem não supervisionada (em reconhecimento de padrões), taxonomias numéricas (em biologia, ecologia), tipologia (em ciências sociais) e partição (em teoria dos grafos) Halkidi (2005).

O problema em torno desse assunto está em como particionar um conjunto em grupos dado um determinado conjunto de dados de tal forma que os dados de um grupo sejam mais semelhantes entre si do que dados em grupos diferentes Guha (1988). Este problema é fundamental pois como trata-se de dados não supervisionados, não se tem a informação de quantas classes existem no conjunto de dados e com isso, não se sabe a quais classes pertencem os objetos ali inseridos.

Dependendo do ponto de vista da análise, um dado pode ser semelhante a outro, porém, mudando a ótica, esses mesmos dados podem ser considerados diferentes, como exemplo disto, pode-se citar a quantidade de grupos. Caso o especialista opte por determinado número de grupos e em uma próxima execução ele altere este parâmetro, os resultados obtidos podem ser totalmente diferentes da primeira execução.

2.1. Algoritmo de agrupamento de dados

Agrupamento de dados pode ser formulado como um problema de otimização com múltiplos objetivos onde o algoritmo a se escolher e seus parâmetros (valores como a função de distância, o limiar de densidade ou o número esperado de clusters) dependem dos dados e do tipo de resultado procurado. Além de o próprio algoritmo poder ser considerado um parâmetro, já que dificilmente sabemos de antemão qual apresentará os melhores resultados.

Quanto à classificação dos algoritmos, pode-se citar a proposta por Jain et al. (1999). Nela, os algoritmos são classificados de acordo com o método adotado para definir os clusters, sendo eles: hierárquicos, particionais, baseados em grid e baseados em densidade.

Um algoritmo de agrupamento hierárquico gera, a partir de uma matriz de proximidade, uma sequência de partições aninhadas. Para formar os agrupamentos, o algoritmo considera uma das alternativas de distância/similaridade entre grupos Faceli et al. (2011).

Este tipo de agrupamento é dividido em duas abordagens: a aglomerativa e a divisiva. A abordagem aglomerativa, começa criando grupos a partir de elementos isolados e vai em direção da formação de um grande grupo contendo todo o conjunto de dados. Já a divisiva começa com todos os elementos formando um único grupo, de modo que divisões sucessivas são feitas até que resulte em elementos isolados.

Já os algoritmos de dados particionais, dividem um conjunto de dados em um

número apropriado de subconjuntos. É uma técnica amplamente utilizada em diversas aplicações, devido a sua simplicidade, facilidade de implementação e rápida convergência.

Um exemplo de algoritmo particional utilizado é o K-means. O algoritmo realiza uma busca para determinar um ponto que represente cada uma das partições, esse ponto é o centro de massa da partição (centroide). Após a definição dos centroides, os pontos são comparados com o centroide e são agrupados com a partição de maior similaridade. Depois de todos os pontos estarem agrupados, o centroide é recalculado e o processo se repete até atingir determinado critério de parada.

Os métodos baseados em densidade permitem descobrir grupos de formatos arbitrários. Estes métodos consideram grupos como sendo regiões densas de objetos no espaço de dados que são separados por regiões de baixa densidade e geralmente representam ruídos Han e Kamber (2006).

DBScan é um exemplo de algoritmo baseado em conceito de densidade tradicional que leva em consideração o centro. Possui como parâmetros de entrada o raio (distância entre um objeto e seus vizinhos) e a quantidade mínima de objetos pertencentes a um determinado raio. A densidade de um objeto é a quantidade de objetos em uma região de alcance, por esse motivo, a densidade de um objeto depende do raio especificado.

3. Avaliação

A avaliação de algoritmo de agrupamento de dados é uma análise não supervisionada de dados. Por este motivo, não pode passar por medidas de erro de estimativas associadas a quaisquer valores observados que é o que ocorre com a classificação, por exemplo, que se trata de análise supervisionada.

Em geral, medição é o processo pelo qual os números ou símbolos são atribuídos a atributos de entidades do mundo real, de tal modo que se possa descrevê-los de acordo com regras bem definidas Fenton e Pfleeger (1996). A medição é um dos passos iniciais para começar um processo avaliativo e isto pode ser aplicado também aos algoritmos de agrupamento de dados.

Porém, o fato de medir algo, muitas vezes não traz os resultados esperados e por esse motivo, para apresentar essas medidas de forma coerente e que possam contribuir para uma análise ou avaliação são necessárias métricas que expressem esses resultados. Com o resultado de métricas podemos melhorar determinados processos ou avaliar como recursos, produtos dentre outros, estão ocorrendo.

Métricas para avaliação de algoritmo de agrupamento de dados não são comuns, porém, este artigo traz algumas delas, que estão abaixo especificadas:

Silhueta: foi proposta por Rousseeuw (1987) e determina a qualidade das soluções com base na proximidade entre os objetos de determinado grupo e na distância desses objetos ao grupo mais próximo. O índice silhueta é calculado para cada objeto, sendo possível identificar se o objeto está alocado ao grupo mais adequado. Esse índice combina as ideias de coesão e de separação Semann (2012). Os valores positivos de silhueta indicam que o objeto está bem localizado em seu grupo, enquanto valores negativos indicam que o objeto está mais próximo de outro(s) grupo(s).

Homogeneidade: A análise de agrupamento tem por finalidade reunir, por algum critério de classificação as unidades amostrais em grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos Johnsosn e Wichern (1992). A métrica referente à homogeneidade tem um resultado satisfatório de agrupamento se todos os seus aglomerados contêm apenas os dados que são membros de uma única classe, ou seja, ela verifica se todos os objetos pertencem àquele grupo e se são homogêneos.

Índice Rand: avalia a concordância de um particionamento com os dados de um particionamento original. O índice verifica sempre um par de grupos e realiza a comparação da similaridade entre os dois. Quanto maior o índice, maior a concordância Rand (1971).

4. GQM (Goal, Question, Metric)

GQM (Goal, Question, Metric) é uma abordagem top-down que visa estabelecer um sistema de medição direcionado a metas. O método foi inicialmente moldado para a área de desenvolvimento de software mas, posteriormente tornou-se parâmetro para muitas iniciativas de medição por ser um meio adequado para alcançar os dados empíricos confiáveis e conhecimento sobre as práticas de software para conduzir possíveis melhorias e um processo sistemático Silva et al. (2009).

A abordagem GQM é baseada na suposição de que para medir algo de forma proposital, deve-se primeiro especificar as metas para si mesmo e implementar isso nos seus projetos, então deve-se traçar objetivos com os dados que se destinam a definir as metas operacionais e, finalmente fornecer uma forma para interpretar os dados com relação às metas estabelecidas Basili et al. (2002).

O GQM é formado por 4 fases. Primeiramente um planejamento é feito, depois a definição de objetivos, questões e métricas e depois inicia a fase de interpretação, nota-se que a fase de coleta de dados se inicia juntamente com a definição e só termina na interpretação pois durante esses dois processos é necessário buscar informações reais e mesmo na fase inicial de planejamento essa busca é feita pois o projeto a ser trabalhado também consta de dados reais.

Para um bom resultado, o objetivo da medição deve estar claro e estruturado. Segundo Basili et al (2002), no objetivo deve estar especificado o propósito (objeto e porque), a perspectiva (qual aspecto e para quem interessa esta informação) e contexto onde está inserido assim como demonstrado na Tabela 1.

Tabela 1: Modelo de definição de objetivos GQM Basili (2002).

Analisar	O objeto a ser medido
Com o propósito de	Entender, controlar ou melhorar o objeto
No que diz respeito ao	Foco da medição
No ponto de vista de	Pessoas que tem interesse na medida
No contexto	Em que a medição tem importância

Depois de estabelecido o objetivo, é a vez de elencar as questões a serem respondidas, as medidas utilizadas e as métricas. A estrutura com objetivo, perguntas, medidas e métricas está representada na Tabela 2.

Tabela 2: Modelo do plano GQM.

Objetivo	Objeto/ Propósito/ Questão /Ponto de Vista/ Contexto.
Pergunta	Questão relacionada ao objetivo.
Medida	Sistema de medida aplicado.
Métrica	Definição da métrica.

O resultado da aplicação do método GQM é a especificação de um sistema de medição visando um conjunto particular de problemas e um conjunto de regras para a interpretação dos dados de medição Solingen e Berghout (1999). Desta forma pode-se analisar cada objetivo e verificar se as métricas escolhidas respondem às questões.

5. Abordagem avaliativa utilizando GQM

A ideia básica de GQM é derivar métricas de software a partir de perguntas e objetivos. Entretanto, embora se originou como uma metodologia de medição para o desenvolvimento de software, os conceitos básicos de GQM podem ser usados em diversas situações em que métricas são necessárias para avaliar satisfação de metas Silva et al. (2009).

A avaliação de algoritmos de agrupamento de dados é uma questão desafiadora. Algumas métricas têm sido relatadas na literatura, porém elas ainda não apresentam resultados satisfatórios. Métodos de avaliação existentes geralmente comprimem os resultados da avaliação em um único número e discordam frequentemente um com o outro, por razões que não são bem compreendidas Shtern (2009).

A utilização de métricas em algoritmos pode trazer a oportunidade de controlar a execução do algoritmo e possivelmente prever qual será sua performance em determinadas situações com a definição de diferentes parâmetros. A medição orientada a objetivos é usada especialmente para programas de melhoramento e esse foi o principal motivo da escolha da abordagem GQM para o propósito do trabalho.

A ideia é propor um modelo de avaliação de algoritmo de dados utilizando-se de métricas existentes na literatura para este fim e utilizar-se do método GQM para localizar, por meio de identificação de objetivos, métricas que se aplicam a determinadas situações, podendo assim, avaliar vários aspectos de um determinado algoritmo.

Solingen (2002), propõe 4 fases do método GQM que foram apresentadas na Seção 5. Baseadas nessas fases, foram definidas 5 etapas que compõe o modelo de avaliação do algoritmo.

Definição: nesta etapa, deve ser definido o problema e o objetivo.

Especificação: formular questões baseadas nos objetivos especificados na etapa de definição;

Medição: estabelecer as métricas que devem responder cada uma das questões feitas.

Demonstração: execução do algoritmo e elaboração de um quadro constando das métricas aplicadas ao problema e comparações entre variação de parâmetros;

Análise: analisar os resultados e elaborar um parecer sobre eles.

Primeiramente devem ser definidos os aspectos básicos da avaliação, a base de dados a ser utilizada, o problema a ser solucionado, qual o algoritmo vai ser implementado para a solução de agrupamento e toda a estruturação do objetivo. Isso é feito na etapa de definição. Na etapa de especificação o objetivo será diluído e as questões referentes a ele devem ser propostas. Depois de definidas as perguntas se dá

início a fase de elencar as métricas que serão utilizadas para responder às questões propostas. Dessas três primeiras etapas resulta o Plano GQM, constando do problema de agrupamento de dados que está sendo tratado, qual o algoritmo que será utilizado e todas as informações à cerca dos objetivos, questões e métricas.

Depois de ter o plano GQM documentado é que se inicia a etapa de aplicação, onde o algoritmo será executado repetidas vezes alterando os parâmetros de configuração do mesmo. Os resultados alcançados com a mudança de parâmetros devem ser relatados em um quadro comparativo. Esse quadro é o resultado da fase de demonstração e se encontra especificado na Tabela 3.

Tabela 3: Modelo para tabela de parâmetros.

Inicialização	Métrica A	Métrica B	...	Métrica N
Parâmetro A	-	-	...	-
Parâmetro B	-	-	...	-
...	-	-	...	-
Parâmetro N	-	-	...	-

E, finalmente, chega-se à etapa de análise, onde o documento de demonstração deverá ser verificado e, baseado nele, uma análise é feita. O modelo da estrutura de avaliação proposta neste artigo e aqui explicado está ilustrado na Figura 4.

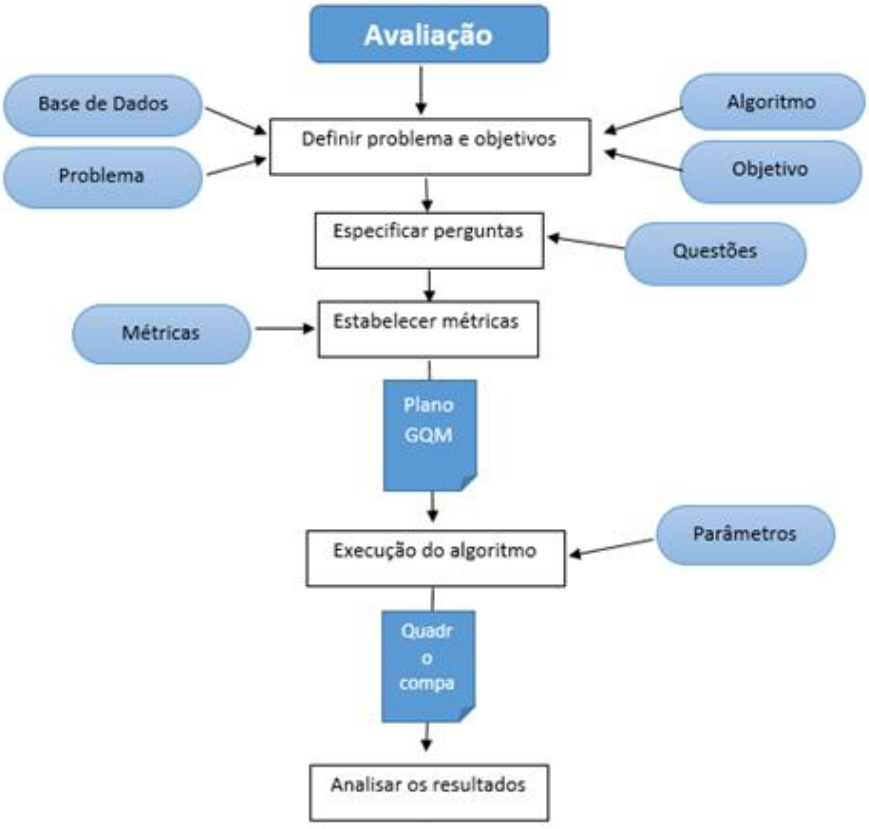

Figura 4: Estrutura do modelo de avaliação de algoritmo baseado em GQM.

A abordagem aqui apresentada tem o propósito de ser simples, porém de representar de forma eficaz métodos para propor objetivos a serem alcançados pelos algoritmos em questão. Podendo assim, chegar à um modelo de avaliação e trazer algumas respostas sobre qual a melhor forma de agrupar determinado conjunto de dados.

6. Teste de Aplicação

Na área de análise de agrupamentos existem várias questões que são consideradas importantes para que um grupo seja bem definido. Os grupos formados podem ser analisados de acordo com a similaridade entre eles, homogeneidade dos elementos pertencentes ao mesmo grupo dentre outros.

Para colocar em prática o modelo proposto para a avaliação de algoritmos de agrupamento de dados utilizando como base o método GQM, um teste de aplicação foi realizado. O estudo aqui realizado é voltado para a avaliação de um conjunto de critérios de configuração em relação a outro conjunto diferente do primeiro, ou seja, as comparações são feitas com um mesmo algoritmo que foi escolhido para ser avaliado em diferentes situações de configuração.

6.1. Dados

O algoritmo escolhido para esta implementação foi o K-means por ser simples e de fácil aplicação e visualização dos resultados. O algoritmo K-means é baseado em centros, ou seja, por meio de diversas iterações, ele adequa os centros e realiza o agrupamento em relação a esses centros encontrados. Por ter essa necessidade de ir readequando os centros ele precisa ser iterado várias vezes e parar quando encontrar o melhor agrupamento para o problema em questão.

O teste aqui apresentado foi realizado com um caso fictício em que a base de dados é gerada automaticamente pelo próprio algoritmo. O algoritmo implementado gera uma base de dados aleatória e a reduz, resultando em 1797 amostras para serem agrupadas e analisadas. Com esse número de amostras, são geradas 10 classes que são utilizadas como parâmetro de configuração em alguns dos conjuntos de parâmetros apresentados.

A saída da presente implementação é um quadro comparativo contendo alguns parâmetros aplicados e as métricas que foram especificadas para o estudo. O algoritmo também plota o resultado do agrupamento em um gráfico para que o usuário possa analisar visualmente a formação dos grupos e seus respectivos centros.

6.2. Plano GQM

Para o estudo apresentado por este artigo, foram estipulados 2 objetivos, um deles busca responder a questão específica do seu desempenho em relação ao tempo de resposta. O segundo objetivo visa verificar a eficiência do algoritmo K-means em relação à qualidade dos grupos formados e para isso, utiliza-se de métricas de homogeneidade, silhueta e índice Rand, vistas na seção 3.1. Os objetivos, bem como suas perguntas, medidas e métricas estão representados no Plano GQM deste experimento Tabela 4.

Tabela 4: Plano GQM para o agrupamento K-means.

Base de Dados	Dados aleatórios (gerados pelo próprio algoritmo).
Problema	Agrupar da melhor forma possível em relação à forma de agrupamento do K-means.
Algoritmo	K-means.
Objetivo 1	**Analisar o algoritmo K-means com o propósito de averiguar o desempenho do processamento no ponto de vista do usuário em relação aos parâmetros de configuração no projeto X do laboratório A.**
Pergunta	Qual o tempo de resposta?

Medida	Contagem de tempo.
Métrica	Tempo de resposta em segundos.
Objetivo 2	**Analisar o algoritmo K-means com o propósito de averiguar a eficiência com que o agrupamento é feito do ponto de vista do usuário em relação aos parâmetros de configuração no projeto X do laboratório A.**
Pergunta	Qual a homogeneidade dos grupos?
Medida	Diâmetro do Grupo e quantidade de elementos
Métrica	Relação entre o diâmetro do grupo e a quantidade de elementos.
Pergunta	Qual a qualidade do agrupamento?
Medida	Medida euclidiana inter-cluster e intra-cluster.
Métrica	Silhueta
Pergunta	Qual a concordância dos dados
Medida	Similaridade entre os grupos (medida euclidiana).
Métrica	Índice Rand

O algoritmo foi implementado duas vezes para este teste, cada uma delas recebeu um conjunto distinto de parâmetros de configuração. A primeira implementação recebeu como parâmetros a Configuração A que é composta por valores distintos para 2 parâmetros considerados principais por muitos autores, esta configuração está demostrada nas Tabela 5 e 6.

Tabela 5: Configuração A.

Parâmetro de Configuração	Valor assumido
Número de clusters	Número de classes (10)
Número de iterações	10

Tabela 6: Configuração B.

Parâmetro de Configuração	Valor assumido
Número de clusters	5
Número de iterações	5

Ambas iterações receberam um conjunto de parâmetros em comum que trata da inicialização dos centros no algoritmo K-means. Foram utilizadas as 3 formas mais conhecidas de inicialização:

K-means++: o algoritmo inicializa os centros dos agrupamentos de forma inteligente, baseado nos pontos de maior homogeneidade de elementos;

Random: inicialização aleatória de centros;

PCA-based: inicializa de forma aleatória e itera o algoritmo apenas uma vez.

A implementação do algoritmo K-means com a Configuração A resultou em números relativamente parecidos para cada métrica aplicada. A métrica baseada no tempo de resposta foi a única que teve diferença significativa em relação à inicialização com PCA-Based, mas esta informação pode ser considerada de baixa relevância, pois o algoritmo é executado apenas uma vez com este tipo de inicialização, eliminando assim as outras 9 iterações que as outras formas de inicialização executaram.

A métrica Silhueta obteve resultados positivos em todas as iterações e com todos os parâmetros, provando que o agrupamento feito pelo algoritmo tem qualidade. A métrica de homogeneidade resultou em todos os casos de inicialização valores em torno de 60% do grupo ser realmente pertencente à classe. O índice Rand apresentou números satisfatórios, porém é preciso lembrar que quanto maior o valor neste caso, maior a concordância dos dados. Os valores resultantes da implementação com a Configuração A podem ser observados na Tabela 7.

Tabela 7: Quadro comparativo da Configuração A.

```
n_digits: 10,    n_samples 1797,
_____________________________________________________________
inicializ   tempo   homog  Rand     Silhueta
k-means++   0.78s   0.602   0.465    0.146
   random   0.70s   0.669   0.553    0.147
PCA-based   0.06s   0.673   0.567    0.150
_____________________________________________________________
```

Já na aplicação do conjunto de parâmetros com a Configuração B, a Silhueta também obteve resultados positivos porém relativamente mais baixos do que na primeira configuração o que teoricamente diminui a qualidade dos grupos formados. A métrica de homogeneidade também teve seus valores reduzidos da mesma forma que o índice Rand. Os resultados da Configuração B podem ser acompanhados averiguando a Tabela 8.

Tabela 8: Quadro comparativo Configuração B.

```
n_digits: 10,    n_samples 1797,
_____________________________________________________________
inicializ   tempo   homog  Rand     Silhueta
k-means++   0.21s   0.364   0.275    0.098
   random   0.27s   0.428   0.350    0.101
PCA-based   0.06s   0.673   0.567    0.130
_____________________________________________________________
```

6.3. Resultados

O teste foi feito aplicando 2 conjuntos distintos de critérios de configuração. O primeiro conjunto (Configuração A) teve a melhor resposta em relação ao objetivo 2 que se refere à eficiência do algoritmo. Pode-se notar que as métricas responsáveis pela eficácia do algoritmo têm seus valores maiores na primeira configuração, isto representa que esta configuração apresenta o melhor agrupamento, uma vez que todas as métricas de eficiência tiveram seus coeficientes aumentados.

Já a Configuração B teve seu melhor desempenho em relação à métrica de desempenho do algoritmo, pois o tempo de resposta do mesmo diminui muito se comparado à Configuração A. Entretanto não se pode dizer que a Configuração B representa o melhor desempenho somente levando em conta esse parâmetro específico, uma vez que o parâmetro relativo ao número de iterações foi diminuído pela metade na segunda configuração.

Neste caso tem-se que desconsiderar o resultado apresentado pela Configuração B em relação ao tempo de resposta e conclui-se que a melhor configuração dentre as duas opções apresentadas no teste é a Configuração A, pois trouxe resultados notadamente melhores para as métricas de eficiência, lembrando que a eficiência de um grupo é uma das características que melhor definem a qualidade.

7. Conclusões e Trabalhos Futuros

Com o presente artigo conclui-se que a dificuldade em avaliar algoritmo de agrupamento de dados não é uma tarefa trivial, pois requer conhecimento aprofundado da área e nota-se que a aplicação dos algoritmos tem formas variadas e em bases de dados variados e com tamanho totalmente diferentes umas das outras o que torna o problema ainda maior.

O método GQM mostrou ser uma opção interessante para a avaliação de algoritmos de agrupamento de dados por ser de fácil implementação e principalmente pela forma como direciona o método avaliativo orientado a objetivos.

GQM é bem definido e suas regras são claras e simples, essa metodologia aliada a intensa busca por métricas de avaliação na literatura tornou clara a busca por métricas que deveriam ser aplicadas para atingir os objetivos propostos. Essa visão foi reforçada pelos resultados obtidos com testes relativamente simples, foi possível demonstrar que com a aplicação do modelo proposto, um algoritmo pode ser avaliado e trazendo resultados visíveis de melhoria em relação a esta ou aquela forma de aplicabilidade.

Posteriormente pretende-se ampliar o modelo com o objetivo de avaliar diferentes algoritmos e dar ao usuário uma forma de optar por este ou aquele algoritmo de agrupamento de dados. Até então esse procedimento é feito exclusivamente baseado no conhecimento que o usuário tem sobre os dados a que serão aplicados o algoritmo e no seu conhecimento na área tornando difícil e muitas vezes errônea essa escolha.

Neste artigo o teste foi voltado para a avaliação de um algoritmo em diferentes situações de configuração de seus parâmetros, o que se pretende no futuro é comparar um algoritmo com outro e avaliar qual dos dois apresenta melhor agrupamento. Isto é de fundamental importância para evolução da análise de agrupamentos, uma vez que hoje não se tem uma base avaliativa para saber qual algoritmo utilizar em determinada situação ficando a cargo do especialista decidir, com base no seu conhecimento e também no problema apresentado, qual ou quais os melhores algoritmos a serem utilizados.

8. Referências Bibliográficas

Basili, V. R.; Caldiera, Gianluigi and Rombach, H. Dieter. (2002) The Goal Question Metric Approach.

Faceli, Katti; Lorena, Ana Carolina; Gama, João e Carvalho André, C. P. L. F de. (2011) Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. (Ed.): LTC. 191-218.

Goldschimidt R.; Passos, E. (2005) Data Mining: um guia prático. Editora Campus, Rio de Janeiro: Elsevier.

Guha, S., Rastogi, R., and Shim K. (1998) CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of the ACM SIGMOD Conference.

Halkidi, Maria et al. (2001) On Clustering Validation Techniques, Department os Informatics, Greece.

Jain, A. K, Murty, M. N. e Flynn, P.J. (1999) Data Clustering: A Review. Unal.

Larose, D. T. (2005) Discovering Knowledge in Data, An Introduction to Data Mining. John Wiley & Sons.

Han, Jiawei e Kamber, Micheline. (2006) Data Mining: Concepts and Techniques. (Ed.): Morgan Kaufamn. Segunda Edição. 2006.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336):846–850.

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

Silva, Carlos V. P. da; Moura, Déborah C. de; Campos, Danylo de C. e Nery, Paulo. (2009) GQM: Goal – Question – Metric.

Shtern, Mark and Tzerpos, Vassilios. (2009) Refining Clustering Evaluation Using Structure Indicators. York University Toronto.

Solingen, Rini van and Berghout, Egon. (1999) The Goal/Question/Metric Method: a practical guide for quality improvement of software development. London. The McGraw-Hill Companies.