Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Colegiado de Engenharia de Computação

Técnicas de Agrupamento para Bases de Dados Grandes

Felipe Moreira Bernardes

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Prof. Eduardo da Silva Ribeiro

Agosto, 2016
João Monlevade/MG

Felipe Moreira Bernardes

Técnicas de Agrupamento para Bases de Dados Grandes

Orientador: Prof. Eduardo da Silva Ribeiro

Monografia apresentada ao curso de Engenharia de Computação do Departamento de Computação e Sistemas da Universidade Federal de Ouro Preto como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação

Universidade Federal de Ouro Preto

João Monlevade

Agosto de 2016

Felipe Moreira Bernardes
Técnicas de Agrupamento para Bases de Dados Grandes/ Felipe Moreira Bernardes. – João Monlevade, 09 de agosto de 2016-
83 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Eduardo da Silva Ribeiro

Monografia (graduação) – Universidade Federal de Ouro Preto, 09 de agosto de 2016.
1. Agrupamento. 2. Grandes Bases de Dados. I. Prof. Eduardo da Silva Ribeiro. II. Universidade Federal de Ouro Preto. III. Instituto de Ciências Exatas e Aplicadas. IV. Técnicas de Agrupamento para Bases de Dados Grandes

CDU 02:141:005.7

Curso de Engenharia da Computação

FOLHA DE APROVAÇÃO DA BANCA EXAMINADORA

Técnicas de Agrupamento para Bases de Dados Grandes

Felipe Moreira Bernardes

Monografia apresentada ao Departamento de Computação e Sistemas da Universidade Federal de Ouro Preto como requisito parcial da disciplina CEA499 – Trabalho de Conclusão de Curso II do curso de Bacharelado em Engenharia da Computação e aprovada pela Banca Examinadora abaixo assinada:

Prof. Me. Eduardo da Silva Ribeiro

DECSI - Universidade Federal de Ouro Preto

Prof. Dr. Matheus Ferreira Satler

DECSI - Universidade Federal de Ouro Preto

Prof. Me. Talles Henrique de Medeiros

DECSI - Universidade Federal de Ouro Preto

João Monlevade, 16 de Agosto de 2016

ATA DE DEFESA

Aos dezesseis dias do mês de Agosto de dois mil e dezesseis, às dezoito horas, na sala C304 do Instituto de Ciências Exatas e Aplicadas, foi realizada a defesa de Monografia pelo aluno **Felipe Moreira Bernardes,** sendo a Comissão Examinadora constituída pelos professores: Prof. Me. Eduardo da Silva Ribeiro, Prof. Matheus Ferreira Satler e Prof. Me. Talles Henrique de Medeiros.

O candidato apresentou a monografia intitulada: *"Técnicas de Agrupamento para Bases de Dados Grandes"*. A comissão examinadora deliberou, por unanimidade, pela aprovação do candidato, concedendo-lhe o prazo de 15 dias para incorporação das alterações sugeridas ao texto final.

Na forma regulamentar, foi lavrada a presente ata que é assinada pelos membros da Comissão Examinadora e pelo graduando.

João Monlevade, 16 de Agosto de 2016

Prof. Me. Eduardo da Silva Ribeiro
Professor Orientador/Presidente

Prof. Dr. Matheus Ferreira Satler
Professor Convidado

Prof. Me. Talles Henrique de Medeiros
Professor Convidado

Felipe Moreira Bernardes
Graduando

TERMO DE RESPONSABILIDADE

Eu, Felipe Moreira Bernardes, CPF: 120.772.616-84, declaro que o texto do trabalho de conclusão de curso intitulado *"Técnicas de Agrupamentos para Bases de Dados Grandes"* é de minha inteira responsabilidade e que não há utilização de texto, material fotográfico, código fonte de programa ou qualquer outro material pertencente a terceiros sem as devidas referências ou consentimento dos respectivos autores.

João Monlevade, 16 de Agosto de 2016

Felipe Moreira Bernardes

Assinatura do aluno

Este trabalho é dedicado à Deus,
o que seria de mim sem a fé que tenho Nele.

Agradecimentos

Primeiramente agradeço a Deus por ter me dado saúde e força para superar todas as dificuldades.

Também à minha família que sempre me apoiou, ajudou em todos os aspectos para que pudesse concluir essa etapa. Aos meus pais Gil e Magna, que me deram todo o suporte, sempre me alegrando com as ligações diárias.

Aos meus irmãos Mateus e Júlia, tios e avós que me apoiavam incondicionalmente. Em especial um agradecimento ao meu tio Rogério, por todas as caronas, conversas e brincadeiras a caminho de casa.

À minha namorada Paula por todo o apoio, que me ajudou em todas as dificuldades, me acompanhou durante esses anos e mesmo longe me fazia sentir calmo e ciente que era capaz de concluir todas as tarefas.

Agradeço aos amigos que fiz nesses anos e que levo por toda a vida, que a distância e o tempo não afastem essa ligação que tivemos, principalmente aos amigos que pude chamar de irmãos da República Bahamas, agradeço por cada momento que passei com cada um deles.

Agradeço também a UFOP que possibilitou todo o conhecimento, ensinamentos e lições não somente educacionais, mas que também agregaram em minha vida e para meu crescimento pessoal. A todos os professores, mestres, técnicos que em sua parte me ajudaram para essa formação. Um grande agradecimento ao Prof. Ms. Eduardo Ribeiro, que me ajudou demais neste trabalho, desde os primeiros dias, me ajudando, sempre que necessário e me passando tranquilidade durante todo o processo.

"Que os vossos esforços desafiem as impossibilidades,
lembrai-vos de que as grandes coisas do homem foram
conquistadas do que parecia impossível."
Charles Chaplin.

Resumo

O crescimento do uso de aplicações, sistemas computacionais e dos usuários vem em uma enorme crescente a cada ano. Os dados gerados por todos tomam proporções gigantescas. Com isso são necessários métodos eficientes, seguros e viáveis para que estes dados possam ser salvos e, utilizados posteriormente de maneira eficaz. Neste trabalho foi realizado o estudo de técnicas de agrupamento com bases de dados grandes, com o objetivo de medir o desempenho destas técnicas e algoritmos, que usualmente são usados em dados menores. As três bases utilizadas apresentavam, respectivamente, dados obtidos de interações reais de usuários com *smartphones*, *smartwatches* e *tablets*. Dados de comerciais de TV de canais indianos e internacionais, tais como fluxo espectral, tempo de saída, etc. E também dados obtidos de vídeos no canal *Youtube* sobre jogos onde usuários discutem sobre os mesmos, tendo como atributos áudio, vídeo e termos relacionados a estes.As bases utilizadas foram obtidas no *UCI Machine Learning Repository*. Os resultados apresentados foram validados utilizando métodos conhecidos na literatura, como o Silhouette, matrizes de afinidades e modelos estatísticos, dando maior respaldo aos dados colhidos. O estudo dessas técnicas, algoritmos e métodos, tanto de validação, descoberta do número ideal de grupos e agrupamento em si, também foi um objetivo deste trabalho. Foram realizados testes em uma plataforma de programação que é propicia para o uso de matrizes, ajudando assim na manipulação de algoritmos utilizados. Ao final do trabalho é possível constatar como o desempenho dos algoritmos apresenta resultados piores, comparados a bases menores.

Palavras-chaves: Agrupamento de Dados. Bases de Dados Grandes. clusters. Aprendizado de Máquina. *UCI Machine Learning Repository*.

Abstract

The growing use of applications, computer systems and users comes in a huge growing every year. The data generated by all take gigantic proportions. With this requires efficient methods, safe and viable so that these data can be saved and later used effectively. In this work the study of clustering techniques with large databases, in order to measure the performance of these techniques and algorithms that are usually used in smaller data. The three bases used showed, respectively, obtained data from real user interactions with smartphones, smartwatches and tablets. Commercial data of Indian and international TV channels such as spectral flow out of time, etc. And data from videos on the channel Youtube on games where users discuss about the same, with the audio attributes, video and related this terms. The data used were obtained from the UCI Machine Learning Repository. The results were validated using methods known in the literature as the Silhouette, affinity matrix and statistical models, providing greater support to collected data. The study of these techniques, algorithms and methods, both validation, ideal number of discovery groups and grouping itself was also a goal of this work. Tests were performed on a programming platform that is favorable for the use of matrices, thereby assisting in the handling of used algorithms. At the end of the work it is possible to see how the performance of the algorithms results are worse, compared to smaller bases.

Key-words: Data Clustering, Clustering Algorithms, Big Datas, clusters, Machine Learning, UCI Machine Learning Repository.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

HHAR	*Heterogeneity Human Activity Recognition*
PAM	*Partitioning Around Medoids*
RAM	*Random Access Memory*
YMVG	*Youtube Multiview Video Games*
WWW	*World Wide Web*

Lista de símbolos

ϕ	Letra grega Phi
θ	Letra grega Theta
σ	Letra grega minúscula sigma
$\in$	Pertence

Sumário

Introdução

As grandes bases de dados são fatores importantes em grandes empresas e serviços dos mais variados tipos. Por isso é necessário estudarmos e entendermos seu funcionamento, tanto teórico quanto prático. É necessário conhecer as técnicas que os envolvem, buscando aperfeiçoar o seu uso. As várias maneiras de organizar e gerenciar esses dados são umas das tarefas que mais se concentram esforços por parte dos pesquisadores. Artigos nessa área estão sendo desenvolvidos para essas técnicas, com o intuito de preparar a tecnologia e desenvolver o conjunto de ferramentas capazes de lidar com tamanho processamento e quantidade de informação. (WU et al., 2014)

Uma dessas vertentes é o agrupamento de dados que pode ser utilizada em diversas áreas e é um dos campos mais estudados. Com ela podemos identificar as semelhanças entre os dados e criar grupos com dados similares, sendo assim, facilitando ao usuário ou algoritmo identificar algum dado específico, que estes buscam. (DUARTE, 2008) A parte de agrupamento se difere dos algoritmos de classificação, principalmente pelos padrões e classificações não-supervisionadas. Além disso, a divisão dos grupos, ou subconjuntos, com coleções não rotuladas de dados significantes, também fazem parte das diferenças. Com uma análise discriminante, podemos obter rótulos dos dados, exclusivamente feita através de uma classificação de padrões pré-classificados como treinamento para que o modelo usado posteriormente aprenda a classificar modelos sem rotulação. (JAIN, 2012)

Ainda segundo (JAIN, 2012) as técnicas de agrupamento nos dão uma forma compacta de representar conjuntos de amostras para obter uma abstração dos dados. Isso é extremamente importante, pois indica uma forma eficiente de processamento, gerando resultados que refletem o conhecimento acerca dos dados, nos dando um processo subjetivo, mas não geral para todos os casos.

Os métodos de agrupamento utilizados devem gerar partições com maior grau de similaridade, buscando um valor, onde essa similaridade seja maior dentro de uma partição do que fora. Ou seja, os grupos devem possuir resultados isolados e com alta coesão interna. Esses resultados devem ser analisados com alguma métrica para expressar essa similaridade, usando distribuições sensíveis, faixa de valores e distribuições espaciais. Com isso, submetemos uma padronização (normalização), de modo a aplicar um pré-processamento a esses resultados, obtendo uma validação com essa técnica.

Motivação

Tendo em vista as técnicas padrões nos conjuntos de dados para se fazer associações, temos várias técnicas capazes de fazer agrupamento de dados, porém todas elas são de valores aproximados. Estudando o uso dessas técnicas em pequenas bases de dados e conseguindo reproduzí-las em uma escala massiva podemos avaliar esse desempenho para que no âmbito das empresas e grandes pesquisas essas técnicas possam ser úteis.

O principal foco do trabalho consiste em analisar o comportamento dos algoritmos e técnicas relacionadas ao agrupamento de dados nas bases de dados grandes (*Big Data*), visto que as bases tendem a possuir maiores valores com o passar do tempo. Verificando como podem ser feitos devido a quantidade massiva de dados, sua escalabilidade, seu modo de agregar em tempo real e constante.

Também verificaremos se os métodos de agrupamento definem uma representação apropriada dos padrões. Tendo em vista uma quantidade vasta de dados, identificando subconjuntos originais, podemos realizar a extração de atributos, e até mesmo computar e compactar novos grupos e/ou atributos.

Objetivos

O objetivo deste trabalho é reconhecer os padrões de dados não rotulados e imprimir técnicas para verificar similaridades e dividí-los em grupos, onde cada grupo possuirá dados semelhantes. Os dados trabalhados são parte de grandes bases de dados, sendo assim podemos dizer que este trabalho tem como foco o estudo das grandes bases de dados, as técnicas de agrupamento e como essa forma de análise é importante no estudo desse campo.

Temos como objetivos específicos:

- Estudar o comportamento e desempenho de algoritmos de agrupamento de dados.
- Estudos comparativos para a validação e qualidade dos agrupamentos gerados.
- Estudar as técnicas de determinação de grupos.
- Comparação dos métodos e testes em grandes bases de dados.
- Analisar os resultados obtidos, para comparação com os testes em outras bases de dados.

Organização dos Tópicos Apresentados

Este texto está dividido em seis capítulos. Em uma introduçao discutimos sobre o problema, abordamos a análise a ser utilizada, tais como as ferramentas propostas para a solução do mesmo. Além disso, damos a motivação para a realização do trabalho, e os objetivos que buscamos alcançar ao fim do texto. No referencial teórico damos atenção sobre as bases de dados, e mostrando quais os dados iremos utilizar no trabalho. Nas seções dois e três , introduziremos, respectivamente, os agrupamentos e seus algoritmos, do que se tratam, e o modo como desenvolvemos e utilizamos suas técnicas. No quarto capítulo é falado sobre os métodos de validação existentes, os mais utilizados, e discursado mais detalhadamente sobre o silhouette, que foi o método utilizado neste trabalho. No quinto capítulo estão as métricas de avaliação que serão usadas para a análise dos resultados obtidos. Na seção dois estarão os testes e os resultados que obtivemos e as respostas que conseguimos extrair dos experimentos. Na última seção estão as conclusões a que chegamos e apontamos sugestões para trabalhos futuros.

Parte I

Referenciais teóricos

1 Grandes Bases de Dados

Estamos na era das grandes bases de dados (*Big Data*). A cada segundo estamos recebendo e armazenando mais informações. Com isso é cada vez mais importantes termos mecanismos capazes de guardar todos esses dados e também fazer com que estejam acessíveis para usuários, pois somente assim estes serão úteis. A implantação destas bases de dados vem desde que temos avanços na tecnologia computacional, com crescimento de aplicações, serviços de *streaming*, lojas virtuais, contas, redes sociais. Cada qual com sua parte de informações, cadastros, formulários, senhas e conhecimento. (WU et al., 2014)

Um sistema que nos permite gerenciar os dados são os Banco de Dados, que é uma coleção de dados inter-relacionados, além de algum programa para acessar esses dados. Esses dados possuem informações relevantes para seus utilizadores cujo principal objetivo dos bancos de dados é armazenar e recuperar essas informações de maneira conveniente e eficiente. Ao serem projetados para lidar com grandes blocos de informação devem definir estruturas complexas de armazenamento e mecanismos de manipulação dos mesmos. (ROUSSEEUW, 1987).

As grandes bases de dados, também chamadas de *Data Intensive Technologies* estão tomando cada vez mais espaço nos campos científicos, industriais e comerciais. Aplicações de empresas como a *Google*, *Flickr*, *Facebook*, possuem inúmeros servidores, nas chamadas "fazendas" destinadas a armazenar e processar os milhares de dados que recebem constantemente. (WU et al., 2014) De fato, podemos dizer que estão presentes em quase todas as atividades humanas, seja para pesquisa, recordações, produção, serviços digitais, etc. (DEMCHENKO; LAAT; MEMBREY, 2014)

Podemos definir as grandes bases de dados usando os 5 V's (DEMCHENKO; LAAT; MEMBREY, 2014):

- **Velocidade:** Devem ser eficientes para trabalhar com os dados, obtendo bom tempo de resposta;
- **Volume:** Devem trabalhar com quantidades gigantescas de dados;
- **Variedade:** Trabalhar com diferentes tipos de dados e garantir entradas e saídas diversas;
- **Valores:** Todos os dados devem possuir valores para que possam ser controlados;
- **Veracidade:** Garantir que os dados sejam verdadeiros e confiáveis;

Ainda segundo (DEMCHENKO; LAAT; MEMBREY, 2014) o que temos que perceber é que todos os processos para se trabalhar com dados, eventos e processos se tornou digital, desde o armazenamento, controle, monitoramento e qualquer outra variável que possa estar contida, a produção de todo conteúdo também é digital. Os métodos de estocagem, a coleção dos dados, classificação e indexação, análise classificação e agrupamento deve ser automática, onde os dados globais devem estar acessíveis e disponíveis para grupos previamente credenciados, pesquisadores, cientistas e em alguns casos acesso público pela internet. Aliado a isso devemos ter uma boa estrutura e ferramentas gerenciadoras que vão agilizar serviços e adaptações que possam ser efetuadas, além de garantir uma segurança aos dados e seus usuários. Para que possam utilizar os serviços de maneira onde não corram riscos de perder dados e ter os mesmos extraídos sem autorização criando um ambiente seguro e confiável a todos.

Com tamanho montante de informações podemos nos sentir imponentes em relação aos dados, tendo um tipo de relação assimétrica, porém devemos explorar novos meios de explorar os dados contidos nessas grandes bases de dados. (SWAN, 2015) Impor novas técnicas de programação, métodos matemáticos e conceitos capazes de compreender e tornar esse conhecimento útil e uma dessas técnicas são os agrupamentos de dados.

Neste trabalho utilizamos as bases de dados que serão listadas a seguir. Estas bases de dados apesar de não possuírem as extensões e quantidade de dados das empresas que foram citadas acima, possuem dados com muitas instâncias, atributos e dimensões, sendo assim podem se passar por grandes bases de dados, se tornando ideais para ajudar em nosso estudo. Todas as bases de dado trabalhadas foram retiradas do repositório *online* UCI *Machine Learnning Repository*. Este contém diversas bases para estudo, aprimoramento de técnicas e algumas onde se pode colaborar com dados.

1.1 HHAR

Conjunto de dados Heterogêneos para Reconhecimento de Atividade Humana, é um conjunto de bases dados de celulares *smartphones* e *smartwatches* (relógios "inteligentes") composto por duas grandes bases de dados onde temos dados de dois sensores presentes na maioria destes dispositivos, o acelerômetro e o giroscópio. Os dados armazenados somam aproximadamente 44 milhões de entradas, e seus atributos são as coordenadas no espaço de uso, modelo do aparelho, usuários e o grau de usabilidade com o aparelho. Além de indexação e valores de log, onde estes dados foram colhidos por pesquisadores de diferentes universidades na Irlanda, Dinamarca e EUA. (STISEN et al., 2015)

Detalhando melhor esta base de dados, está composta por treze modelos diferentes de quatro fabricantes. Os sensores, dispositivos e carga de trabalho compõem todos os dados que são divididos em duas grandes tabelas onde essas tabelas possuem 11 colunas

de atributos, com aproximadamente três milhões e seiscentas mil linhas, onde cada linha indica um dado colhido. Cada coluna tem seu significado, sendo que os mais relevantes para a composição dos dados são o modelo do dispositivo, sua categoria para a pesquisa de coleta (STISEN et al., 2015), e as coordenadas colhidas para os sensores de acelerômetro e giroscópio.

Tabela 1: Características da Base de Dados HHAR

Características dos Dados	Multivariados, Série de Tempo	**Número de Instâncias**	43930257	**Área**	Computação
Atributos Característicos	Real	**Número de Atributos**	16	**Data de Doação**	25/10/2015
Tarefas Associadas	Classificação, Agrupamento	**Faltando Valores ?**	Sim	**Número de Visitas WEB**	24590

1.2 TV News Channel Commercial Detection Dataset

Este conjunto de dados também foi retirado do repositório UCI de bases de dados para aprendizado de máquina e é o resultado de monitoramento automático para identificar blocos comerciais em vídeos de notícias, e aplicações no domínio de análise de transmissões televisivas. Estes dados consistem em comerciais que ocupam de 40 a 60 por cento do tempo total de transmissão de emissoras de televisão, e com isso esses dados foram colhidos com o intuito de categorizar e agrupar estes comerciais exclusivamente pela apresentação audiovisual. (VYAS et al., 2014)

Esta base de dados é ideal para trabalhos com aprendizado de máquina, pois os dados colhidos são de emissoras de TV indianas, nacionais e internacionais, que não possuem nenhum formato de apresentação, e assim, tendem a possuir grande variabilidade e natureza dinâmica. Neste estão cento e cinquenta horas de transmissão de cinco canais de maior audiência. São eles: CNNIBN, NDTV, TIMESNOW, BBC e CNN. As gravações estão em resolução 720x576 a 25 fps. (VYAS et al., 2014)

As filmagens são utilizadas como instância geral. Alguns atributos importantes presentes são o histograma RGB entre frames consecutivos, os diferentes tipos de áudio captados, nesse caso sendo sete, taxa de cruzamento zero, centroide espectral, fluxo espectral, faixa de frequência, frequência fundamental, palavras encontradas no áudio e tipos de vídeo, nesse caso cinco ao todo.

O entendimento desses atributos neste trabalho fica somente focado em graus de correlação, entre os mesmos, pois estávamos interessados em seu grau de agrupamento, similaridade e tempo de execução nos algoritmos.

Na tabela a seguir as características dessa base de dados segundo a UCI:

Tabela 2: Características da Base de Dados TV News

Características dos Dados	Multivariados	**Número de Instâncias**	1029685	**Área**	Computação
Atributos Característicos	Real	**Número de Atributos**	12	**Data de Doação**	27/03/2015
Tarefas Associadas	Classificação, Agrupamento	**Faltando Valores ?**	Não informado	**Número de Visitas WEB**	19460

1.3 Youtube Multiview Video Games Dataset

A base de dados YMVG foi coletada por Madani Omid (MADANI; GEORG; ROSS, 2013) em parceria com o time de análise e percepção de vídeos da Google. É uma base utilizada para testes com classificação e agrupamento disponibilizada pelos autores na UCI Machine Learning.

Consiste de valores e classes de rótulos para cerca de um milhão e duzentas mil instancias (vídeos) onde cada uma dessas é descrita com diferentes tipos de atributos. Tendo três dimensões principais: texto, visual e áudio. Apesar de não utilizarmos os rótulos, existem trinta tipos diferentes, cada um sendo de algum vídeo game popular.

Nenhum nome destes jogos foi mencionado, portanto os dados são puramente numéricos. Segundo(MADANI; GEORG; ROSS, 2013), estes dados são bastante uteis para uso em pesquisas de aprendizado multi-níveis, incluindo o caso deste trabalho, aprendizado de agrupamento, supervisionado e não-supervisionado. propício para testes com diferentes técnicas e algoritmos, o que explica o seu uso para o presente trabalho.

Atualmente, algumas pessoas ao jogarem algum game, a certo ponto gravam vídeos, e carregam ao *Youtube* (plataforma de vídeos), compartilhando experiências de jogo, conhecimentos, etc. Esses vídeos são essencialmente o que compõem esta base de dados, somando 300 mil dos dados. As classes destes vídeos são importantes para construir uma base de classificação, pois dominam sobre outras classes, ou seja, tem maior correlação. (MADANI; GEORG; ROSS, 2013)

Os dados presentes seguem um formato esparso, indicando linhas onde faltam alguns dados. Este problema é contornado om uma manipulação da base para que esses dados não comprometam os algoritmos, gerando algum tipo de erro.

A seguir as características da base:

Tabela 3: Características da Base de Dados YMVG

Características dos Dados	Multivariado, texto	**Número de Instâncias**	1200000	**Área**	Computação
Atributos Característicos	Real, Inteiro	**Número de Atributos**	31	**Data de Doação**	16/10/2013
Tarefas Associadas	Classificação, Agrupamento	**Faltando Valores ?**	Sim	**Número de Visitas WEB**	53179

2 Agrupamento de Dados

Como vimos, ao trabalhar com dados multivariados é importante que tenhamos uma maneira de organizar os mesmos, trabalhando de forma a facilitar o entendimento do banco de dados. A análise de grupos tem sido bastante utilizada para este tipo de entendimento e possui uma vasta literatura que vem crescendo continuamente e mostra a importância dessa área. (JAIN, 2012)

Podemos chamar o processo de agrupar dados similares de agrupamento. Existem diversas técnicas para realizar este tipo de agrupamento e separar classes de dados distintas. Essas técnicas visam selecionar o melhor agrupamento dentre os possíveis para que se ajuste aos dados relativos.(DUARTE, 2008)

O agrupamento de dados vem sendo amplamente utilizado em diferentes campos, geralmente associados a taxinomias, como na área médica e veterinária. (HAMAD; BIELA, 2008) Outras áreas onde esse modelo é utilizado são:

- **Marketing:** Encontrar grupos de consumidores com um comportamento similar, pode conter uma base de dados grande contendo bens e compras;
- **Livrarias:** Otimizar a organização de livros;
- **Planejamento de Cidades:** Identificar grupos de residências com base em valor, localização, etc.;
- **WWW:** Classificação de documentos, agrupamento de dados em blogs para descobrir padrões de acessos similares;
- **Astronomia:** Identificar diferentes tipos de astros;

Enfim, são diversas áreas de atuação onde o agrupamento de dados pode ser aplicado, sempre com o intuito de facilitar e catalogar os diversos tipos de dados recebidos. Esta é uma ferramenta para se ter a percepção da distribuição dos dados, identificar as características de cada grupo e focalizar em um conjunto de grupos para análise. Pode também ser usada como um passo de pré-processamento para outros algoritmos, além de ser usada para compreensão dos dados pelo uso dos objetos representativos do mesmo e geração e teste de hipóteses.(LANGONE et al., 2016)

2.1 Tipos de Agrupamento

O agrupamento de dados, muitas vezes referido somente como agrupamento, consiste na divisão de um conjunto de dados em grupos, de forma a colocar objetos de dados semelhantes no mesmo grupo e objetos de dados dissemelhantes em grupos diferentes. (MAXWELL, 2009)

E de acordo com a necessidade da aplicação onde esses dados serão aplicados pode haver interpretações diferentes. Para isso temos alguns modelos de agrupamentos que condizem com determinados tipos de dados. Segundo (MAXWELL, 2009) temos as seguintes definições:

- **Hierárquico:** Neste modelo os grupos de objetos estão organizados como uma árvore.
- **Particionado:** Os dados deste tipo são divididos em grupos de mesmo nível, ou seja, sem sobreposição de clusters ou não- aninhada.
- **Exclusivo:** Neste agrupamento cada dado está correlacionado a apenas um cluster.
- **Sobreposto:** Ao contrário do agrupamento exclusivo, este modelo aceita dados que existam em diferentes clusters simultaneamente.
- **Fuzzy:** No agrupamento *fuzzy* temos um percentual (grau de pertinência) pertencente aos dados, que os relacionam com os clusters advindos.
- **Completo:** Neste modelo os dados necessariamente devem estar associados a um cluster.
- **Parcial:** Diferente do modelo completo alguns dados podem não ser bem definidos e não participarem de nenhum grupo ou cluster.

Devido a serem mais estudados e citados nas pesquisas e trabalhos acadêmicos, além de serem mais simples de serem trabalhados podemos aprofundar nos métodos Hierárquicos. Segundo (LOPES, 2009) as técnicas usadas são simples e o modo de atuação são sucessivos particionamentos, produzindo uma representação hierárquica dos agrupamentos, no caso árvores, o que torna as técnicas utilizadas nesse método simples é a não utilização de um número prioritário de agrupamentos. Este método requer a utilização de uma matriz contendo os valores de distância entre os agrupamentos em todos os estágios do agrupamento, numa matriz que é conhecida como matriz de similaridades. Os métodos Hierárquicos ainda podem ser divididos em métodos divisivos e métodos aglomerativos.

Os métodos divisivos não serão utilizados nem estudados devido a sua ineficiência e por exigir uma grande capacidade computacional. Já os métodos aglomerativos se iniciam

com um padrão formado e gradualmente os grupos são unidos até que um grupo com todos os dados seja formado, assim no começo temos vários grupos com poucos elementos e alto grau de similaridade entre os grupos e ao final do processo temos poucos agrupamentos e com pequeno grau de similaridade e muitos elementos.(COSTA; NETTO, 1999)

Segundo (JAIN, 2012) sobre os métodos Particionados, estes são baseados na minimização de uma função de custo, onde os padrões são agrupados em um número definido de agrupamentos escolhidos anteriormente. Cada padrão é agrupado na classe em que essa função é minimizada.

Os métodos particionados tem vantagens em relação aos métodos hierárquicos, além de serem extremamente mais rápidos. A principal vantagem é a possibilidade de um padrão mudar a medida que o algoritmo evolui em sua execução, do mesmo modo em que estes métodos são capazes de trabalhar com bases de dados grandes, o que é um ótimo método a ser estudado e desenvolvido neste trabalho. Um ponto a ser explorado por esses métodos são exatamente o número de agrupamentos que são escolhidos a priori, pois este número pode acarretar em interpretações erradas caso o valor escolhido não seja o ideal. O algoritmo é sensível às condições impostas e não busca a estrutura inerente ao número de agrupamentos ideal. (LOPES, 2009)

2.2 Tipos de Aprendizado

O termo de aprendizagem de máquina está diretamente ligado ao reconhecimento de padrões, essa sendo a área mais explorada, ou seja, dado um conjunto de treinamento, deseja-se predizer o comportamento dos dados desconhecidos. Contudo na área de aprendizagem automática, vários algoritmos eficientes vêm sendo desenvolvidos para realizar tarefas cada vez mais complexas, e com isso a parte de compreensão teórica também vem sendo melhor explicada. A variável que devemos avaliar ao falarmos de aprendizagem é o conhecimento, prévio ou não, dos dados em questão avaliados e dependendo desta variável a aprendizagem pode ser classificada como supervisionada, não supervisionada ou semi-supervisionada. (LELE; MOUTARI, 2008)

Na aprendizagem supervisionada temos apenas classificação, com dados rotulados, ou seja, o conjunto de dados é pré-classificado, e aprende descrições das classes já existentes do mesmo. Vemos que o objetivo desta técnica é aprender as regras de decisão com dados de treinamento, sendo assim construindo classificadores na fase de aprendizagem e posteriormente usar esses classificadores para determinar os grupos do restante dos dados ainda não classificados.

A combinação de diversos classificadores supervisionados também é considerada um aprendizado supervisionado e esta combinação se mostra vantajosa e eficiente levando a obtenção de melhores resultados do que aplicações isoladas.

Na aprendizagem não supervisionada já temos o agrupamento e trabalhamos apenas com dados não rotulados. É um problema de difícil resolução (ARORA; CHANA, 2014) já que a princípio não se conhece o número de classes existentes nem a qual classe os dados estão associados, para resolvê-lo devemos obter grupos ou classes de dados que possuem similaridades entre si e dissimilaridades aos dados de diferentes grupos. Este tipo de aprendizagem possui alguns parâmetros que podem levar a diferentes estruturas para um mesmo conjunto de dados.

Ainda segundo (ARORA; CHANA, 2014) os algoritmos de aprendizagem semi-supervisionado podem ser descritos como sendo híbridos, já que trabalham com dados rotulados e não rotulados. Podemos ver esse problema como uma classificação supervisionada onde não se conhece todas as classes dos dados de treinamento, para que possamos realizar a classificação é necessário a inserção de conhecimento adicional, relativo a estrutura de dados que se quer obter. Esse conhecimento está na forma de restrições, isto é, descreve-se se determinados tipos de objetos devem ficar no mesmo grupo ou não, com isso identificamos ligações obrigatórias ou proibidas entre estes objetos.

Como dito acima o reconhecimento de padrões é um campo bastante estudado atualmente, e usando das ideias da aprendizagem, verificou-se a combinação de classificadores supervisionados a problemas de aprendizagem não supervisionada. Essa combinação tem como finalidade minimizar problemas inerentes ao processo de agrupamento de dados e criar um agrupamento final que seja melhor do que aquele que lhe deu origem.(ARORA; CHANA, 2014)

3 Algoritmos de Agrupamento

Como vimos, as aplicações de agrupamento de dados têm como requisito de vital importância o tempo de processamento. Sendo assim, é grande a procura por algoritmos eficazes e eficientes para bases de dados que continuam a crescer, por isso que os algoritmos hoje buscam satisfazer requisitos específicos, tais como: escalabilidade, alta dimensionalidade, interoperabilidade e utilidade, etc.

Com o aumento do conjunto de dados analisado, o desempenho do algoritmo não pode diminuir. Hoje em dia os algoritmos, geralmente, trabalham com conjuntos de dados de pequenas dimensões (QUEIROZ, 2009). Seguindo o mesmo aspecto, os algoritmos atuais costumam trabalhar com duas, a três dimensões de grupos de dados, alguns sendo bastante assimétricos em um espaço disperso, porém o ideal seja que os algoritmos mantenham o desempenho à medida que o espaço e número de dimensões cresce de forma abrupta.

A descoberta dos grupos que se diferem é obtida a partir de medidas euclidianas, ou de Manhattan, baseados na forma, tamanho e densidades e com possibilidade de tratar dados de diferentes tipos, na maioria das vezes contínuos. (MAXWELL, 2009)

Como alguns algoritmos necessitam de se introduzir parâmetros, como no caso das medidas citadas acima, alguns resultados ficam sensíveis a esses parâmetros, e isso pode levar a algumas situações que devem ser consideradas para o controle de qualidade do agrupamento. Uma das coisas que pode levar a um agrupamento de baixa qualidade são a robustez e o ruído, visto que, a maioria dos conjuntos de dados contém valores isolados, ou dados com erros e/ou desconhecidos, alguns algoritmos podem ser sensíveis a ordem de entrada dos dados, gerando dados diferentes de acordo com essa ordem. E de modo a facilitar o entendimento, a interoperabilidade dos agrupamentos deve ser de simples interpretação, compreensão e passíveis de serem utilizados, pois na maioria dos casos, estes estão associados a aplicações de semânticas específicas. (QUEIROZ, 2009)

Com essas considerações e pesquisas realizadas é comprovado que nenhum algoritmo de agrupamento atual cobre todos os requisitos corretamente, ou boa parte deles. (CALON et al., 2015)

3.1 K-Médias

Este algoritmo é um método de agrupamento particionado. Usado quando já se tem a hipótese do número de grupos, este número pode ser passado ao algoritmo para que forme a quantidade de agrupamentos fornecida inicialmente. Portanto este método irá gerar exatamente k grupos com a maior distinção possível entre eles. (MAXWELL, 2009)

Segundo (LOPES, 2009) este algoritmo é iterativo e minimiza a soma das distâncias de cada padrão ao centroide de cada agrupamento, movendo os padrões entre os agrupamentos, a função objetivo não se altera, ou tem um mínimo de alteração. O resultado disso é um conjunto de agrupamentos compactos e bem separados.

(MAXWELL, 2009) ainda apresenta um pseudocódigo, ou um conjunto de passos a ser seguido para se construir o algoritmo de K-Médias:

Algoritmo 1: K-MÉDIAS

Entrada: Matriz de dados

Saída: Médias das k partições

1 **inicio**

2 Para cada padrão determinar a partição mais próxima.

3 Calcular a média de cada partição.

4 Se houver mudanças nas médias das partições, retorne passo 2.

5 **fin**

6 **retorna** k

A solução encontrada pelo K-Médias está totalmente relacionada com a escolha das condições iniciais e do mínimo local e apesar de se tratar de um método computacionalmente eficiente, está sensível a ruídos.(ARORA; CHANA, 2014).

Como no trabalho de (NG et al., 2002), o algoritmo de K-Médias pode ser usado para complementar outro método. Isso é feito pois os agrupamentos não correspondem a regiões convexas. O algoritmo K-Médias roda diretamente para encontrar agrupamentos não satisfatórios. (NG et al., 2002) No artigo de Ng, foi utilizado o spectral clustering que abordaremos na próxima seção.

3.2 Kernel K-Médias

O kernel K-médias se difere do homônimo K-médias, quando assumimos um padrão diferente de como realmente estão distribuídos os dados, geralmente ao utilizarmos o K-médias com uma distância Euclidiana, esperamos encontrar dados em regiões elípticas, embora possamos encontrar dados em regiões diferentes. Assim devemos realizar transformações nos dados, para que assim possam ser usados em um novo espaço. (ZHANG; RUDNICKY, 2002)

Então podemos adaptar o método K-médias aplicando uma função *kernel* (núcleo) não linear capaz de mapear os dados de entrada originais para uma dimensão de maior nível, deste modo, podemos extrair agrupamentos não lineares que estão separados do espaço de entrada. (DHILLON; GUAN; KULIS, 2004)

Ainda segundo (DHILLON; GUAN; KULIS, 2004) é possível generalizar os algorit-

mos kernel K-médias introduzindo um peso para cada ponto representado nos dados. Com esse peso e as funções não lineares, podemos construir um algoritmo poderoso. As funções não lineares que são mais utilizadas são mostradas na tabela a seguir:

Tabela 4: Funções kernel

Kernel Polinomial	$k(a,b) = (a.b + c)^2$
Kernel Gaussiano	$k(a,b) = e^{\|\|a-b\|\|/2\sigma^2}$
Kernel Sigmoide	$k(a,b) = tanh(c.(a.b) + \theta)$

Nas funções temos ai e bj como vetores de colunas, ou linhas, dependendo da forma como são dispostos os dados. A variável c é um parâmetro geralmente com o valor 1. E a variável d indica em quantas dimensões estamos trabalhando. Portanto, podemos considerar dados R^d, com d = 1,2,3... E os parâmetros θ e σ são definidos inicialmente, tendo valores fixos durante toda a execução.

Os algoritmos de kernel k-médias e *spectral clustering* podem se relacionar, por meio dos pesos impostos nos algoritmos e das funções de *kernel* aproximadas, para utilizar uma função que gere uma solução maximizada do problema, os pesos devem ser reescritos. Isso nos dá um esquema que se assemelha ao método baseado em *spectral clustering*. (DHILLON; GUAN; KULIS, 2004)

Algoritmo 2: KERNEL K-MÉDIAS

Entrada: K,k, w

Saída: $C_i ... C_k$

1 **inicio**

2 Inicializar grupos $C_i^0, ..., C_k^0$;

3 Inicializar t=0;

4 Para cada a, encontrar grupo com índice $j(a) = argmin_j ||\phi(a) - m_j||^2$, usando distância Euclidiana.

5 Computar os agrupamentos $C_j^{i+1} = a : j(a) = j$

6 Se não convergir setar t para t+1 e voltar ao passo 4. Caso contrário, pare.

7 **fin**

3.3 K-Medóides

Ao contrário do algoritmo K-Médias, este algoritmo considera o valor médio dos objetos como ponto de referência. O dado mais centralmente localizado no grupo se torna o medóide e apesar das diferenças este algoritmo também utiliza o princípio da minimização da função do erro quadrático. (DUARTE, 2008)

Ainda segundo (DUARTE, 2008) a estratégia desse algoritmo consiste em substituir um dos medóides por uma não medóide quando a qualidade do agrupamento resultante é melhorada. Esta qualidade é medida usando a função objetivo ou de custo. Esta medida é a diferença média entre os dados e o medóide do seu grupo. A função de custo calcula a diferença do valor do erro quadrático se um medóide for substituído e cada vez que uma re-atribuição acontece, a diferença no erro quadrático contribui para a função de custo. Se ao final de cada atribuição a função for negativa então um novo medóide é selecionado. Caso a função retorne positiva então esse medóide é aceito pelo algoritmo. O agrupamento resultante será um conjunto K de grupos que minimizam a soma das dissimilaridades de todos os objetos ao medóide mais próximo.

Um dos primeiros e mais robustos algoritmos K-Medóides foi o PAM (*Partitioning Around Medoids*)que contém os padrões onde a dissimilaridade média entre os pertencentes a um certo agrupamento é mínima. Além de minimizar a soma das dissimilaridades, este algoritmo não depende de uma suposição inicial para o centro dos agrupamentos. (MAXWELL, 2009)

Algoritmo 3: K-MEDÓIDES

Entrada: X, n

Saída: P

1 **inicio**
2 Escolher arbritariamente k objetos de dados X com os medóides iniciais
3 Atribuir objetos não medóides ao grupo com medóide mais próximo.
4 Selecionar aleatoriamente objeto não medóide
5 Calcular custo total S.
6 Se S <0 troca-se o medóide para formar um novo conjunto de k medóides.
7 Até que não haja mais mudanças
8 Define-se P como o agrupamento com os grupos obtidos no passo anterior
9 **fin**

3.4 Fuzzy c-Médias

Esta é uma técnica de agrupamento de dados *fuzzy*. É uma evolução das técnicas de agrupamento de dados mais recentes e fornece um método que mostra como agrupar padrões que pertencem a um espaço multidimensional em um número específico de agrupamentos. (MAXWELL, 2009)

Primeiramente neste método temos uma suposição inicial sobre os centros de cada agrupamento, para cada padrão é assinalado um grau de pertinência para cada agrupamento e esse padrão é atualizado iterativamente, assim como os centros. Essa iteração tem como objetivo minimizar sua função objetivo que representa a distância de

cada padrão em relação ao centro de cada agrupamento. (LOPES, 2009)

Ainda segundo (LOPES, 2009) pseudo-partições fuzzy são criados para divisão dos padrões de um conjunto de grupos nos dados e para isso existe um coeficiente chamado coeficiente *fuzzy*, um número real, que controla a influência dos graus de pertinência nessas partições. O autor mostra em seu artigo que aproximando o coeficiente (m), $m \longrightarrow 1$ então o algoritmo tem um comportamento semelhante ao K-Médias. Quando $m \longrightarrow \infty$ então os centros do agrupamento ficam mais próximos do centroide do conjunto de dados e a variância de cada agrupamento se torna maior, tornando os agrupamentos mais *fuzzy*.

Algoritmo 4: FUZZY C-MÉDIAS

Entrada: X, c, m, ε

Saída: P

1 **inicio**

2 Selecionar partição pseudo Fuzzy e centros dos clusters;

3 Calcular centros dos clusters para pseudo partição;

4 Atualizar pseudo-partição Fuzzy;

5 Se diferença entre pseudo-partição (t) e pseudo-partição $(t+1)$ for menor ou igual ao erro ε lança erro

6 Senão retorna pseudo-partições criadas e seus centros;

7 **fin**

3.5 Spectral Clustering

Este método de agrupamento é bastante poderoso usando um grafo particionado dos dados e utiliza-se uma função objeto para encontrar a partição ótima do grafo. Essa partição ótima corresponde a uma solução de valores próprios, o agrupamento espectral usa esta solução ótima para gerar o resultado dos agrupamentos finais para o conjunto de dados fornecidos.(SHINNOU; SASAKI, 2008)

Segundo (HAMAD; BIELA, 2008) existem algumas maneiras de se criar este grafo. Para representar a relação entre os dados podemos construir:

- **Grafo totalmente conectado:** Todos os vértices com similaridades não nulas estão conectados.
- **Grafo R-vizinhos:** Cada vértice dentro de um raio R é ligado. R é um valor real que tem de ser ajustado a fim de capturar a estrutura de dados.
- **Grafo de vizinhos k-próximos:** Cada vértice é conectado com o vizinho K mais próximo. K é um valor inteiro que controla a relação de dados local.

Neste método estamos considerando construir um grafo de similaridades. O objetivo da construção desse grado é modelar a relação de vizinhança entre os objetos de dados. Segundo (DUARTE, 2008) este algoritmo transforma o problema do agrupamento de dados em um problema de otimização combinatória.

O conjunto de dados agora pode ser representado como um grafo G (V, A), onde (V) são os nós e (A) são as arestas. Uma aresta que liga quaisquer dois vértices tem uma ponderação igual a uma medida de similaridade entre os dados. Um conjunto de arestas cuja remoção divide o grafo em K subgrafos é denominado separador de arestas. O objetivo do algoritmo é encontrar um separador com uma soma mínima das ponderações das arestas eliminadas, isto é, encontrar o corte mínimo. (DUARTE, 2008)

As ferramentas principais do agrupamento espectral são as matrizes de *Laplacian*, porém como (DUARTE, 2008) explica, não existe uma convenção única de qual matriz *Laplacian* exatamente utilizar. Cada autor usa a "própria" matriz, sendo as duas principais: normalizada e não-normalizada.

No trabalho de(SHINNOU; SASAKI, 2008) o método de agrupamento espectral não obtém bons resultados para uma base de dados grandes, porém os autores ressaltam que a adição de algoritmos como o K-Médias podem reduzir o problema em matrizes de convergência e similaridade menores, a ponto de conseguir um desempenho eficiente para uma base de dados independentemente do tamanho e diversidade dos dados.

Além disso a matriz de similaridade não deve ser esparsa, pois assim não sofre do problema do mínimo local. Não é necessário reinicializar o algoritmo várias vezes e com diferentes inicializações, e a desvantagem deste método é a instabilidade, e a não trivialidade da escolha dos parâmetros de vizinhança na construção do grafo, além da

escolha da matriz de similaridade.

Algoritmo 5: SPECTRAL CLUSTERING

Entrada: S,k

Saída: Y

inicio

Formar matriz de afinidade $A \in R^{nxm}$, com $A_{ij} = e^{(-||s_i - s_j||^2/s\sigma^2)}$, se $i \neq j$ então A = 0;

Definir matriz diagonal cuja i-ésima lina de A construa a matrz $L = D^{-0.5}.AD^{-0.5}$

Achar $x_i, ..., x_k$ dos k maiores auto vetores de L e formar matriz $X = x_i, ..., x_k \in R^{nxk}$, empilhando os autovetores nas colunas;

Formar matriz Y normalizando linhas de X para um tamanho único

Tratar cada linha de Y em cada pont R^k, em k clusters com outro algoritmo de agrupamento

Designar o ponto originalS_i para gtupo j somente se linha Y for pertencente ao grupo;

fin

4 Modelos de Validação

A validação de agrupamentos de dados, junto com a descoberta do número apropriado de grupos são temas bastante estudados e temos por objetivo selecionar dentre os agrupamentos criados, o que melhor se ajusta aos dados fornecidos. Essa validação pode ainda ser baseada em critérios internos ou externos, de modo que capture a maioria das propriedades de um bom agrupamento. (DUARTE, 2008)

Para obtermos um agrupamento de dados que melhor se adapta aos respectivos dados utilizados, devemos utilizar de técnicas de validação. Usando de medidas para que as propriedades desejadas satisfaçam o algoritmo, na maioria dos índices, as propriedades buscadas são a dispersão e a separação. (DUARTE, 2008)

Geralmente, segundo (LOPES, 2009), as propriedades consistem na análise dos dados quantitativos presentes nos resultados, usados de forma objetiva gerando uma quantidade de agrupamentos expressos em um número real. Ainda indicando o grau das propriedades determinadas no mesmo.

Pode-se usar dois casos diferentes, em um não temos o número K de grupos como parâmetro, e outro onde temos este valor. No primeiro executamos os algoritmos de agrupamento em um intervalo de valores, e temos como resposta uma média dos valores onde é determinado os melhores valores para o conjunto de dados, além de identificar o número de grupos do mesmo. No segundo caso é usado índices de validação para que obtenhamos o melhor valor de K. Executamos o algoritmo para valores previamente determinados, onde para cada valor o algoritmo é executado um número r de vezes com valores de parâmetros diferentes, e após isso, construímos gráficos representando os índices de validação para cada valor de K considerado inicialmente com base na continuidade do gráfico consideramos os valores máximo ou mínimo da curva como o melhor valor de agrupamento. (DUARTE, 2008)

4.1 Silhouette

O método silhouette interpreta e valida a consistência de um agrupamento de dados. É estimado a média entre as distâncias entre os agrupamentos criados. Basicamente o método vai avaliar quão similar os dados de um mesmo agrupamento são, e qual a média de diferença com os demais agrupamentos. Mensurando valores entre -1 e 1, onde os menores valores indicam maior grau de separação e os maiores valores representam um maior grau de coesão.

Com essas afirmações podemos dizer que um valor médio alto nos dá uma configu-

ração de agrupamentos apropriados.

Tendo aplicado as técnicas de agrupamento, basicamente precisamos de dois dados para construir o método silhouette e a partição obtida é uma delas, a coleção de valores de proximidade entre os dados.

Matematicamente, segundo (CHARRAD, 2016), podemos dizer que:

Temos valores a(i) sendo a medida classificada para cada iteração i de dados. Tendo b(i) como o menor valor calculado de dissimilaridade com os demais agrupamentos no qual i não faz parte. O valor de b(i) pode ser descrito como o agrupamento vizinho mais próximo. Assim podemos definir o método silhouette como:

$$s(i) = \frac{a(i) - b(i)}{max[a(i), b(i)]}$$

onde s(i) é o resultado do método para cada dado i.

Com isso verificamos que o valor de s(i) é: -1 <= s(i) <= 1.

Com estas definições podemos tirar algumas conclusões acerca dos dados de entrada. Como, por exemplo, para obtermos um valor próximo de 1, ou seja, ou valor coeso, precisamos de uma média a(i) bem menor que b(i). Outro ponto que pode ser tratado é quando o valor de s(i) é próximo de 0, isso significa que os dados estão próximos da fronteira entre os agrupamentos existentes. (ROUSSEEUW, 1987)

Podemos plotar o gráfico com cada dado i, combinado com a média de s(i), para cada teste de k agrupamentos, como visto na Figura 1. Nesse gráfico conseguimos identificar quais dados estão dentro de seus determinados particionamentos, e quais são meramente figurados entre os grupos criados. Todo o agrupamento é a combinação através do silhouette em uma única parcela. Isso permite uma avaliação da qualidade dos grupos e uma visão da configuração geral dos dados assim dispostos. Utilizando portanto o silhouette, selecionamos o número apropriado de agrupamentos a serem criados.

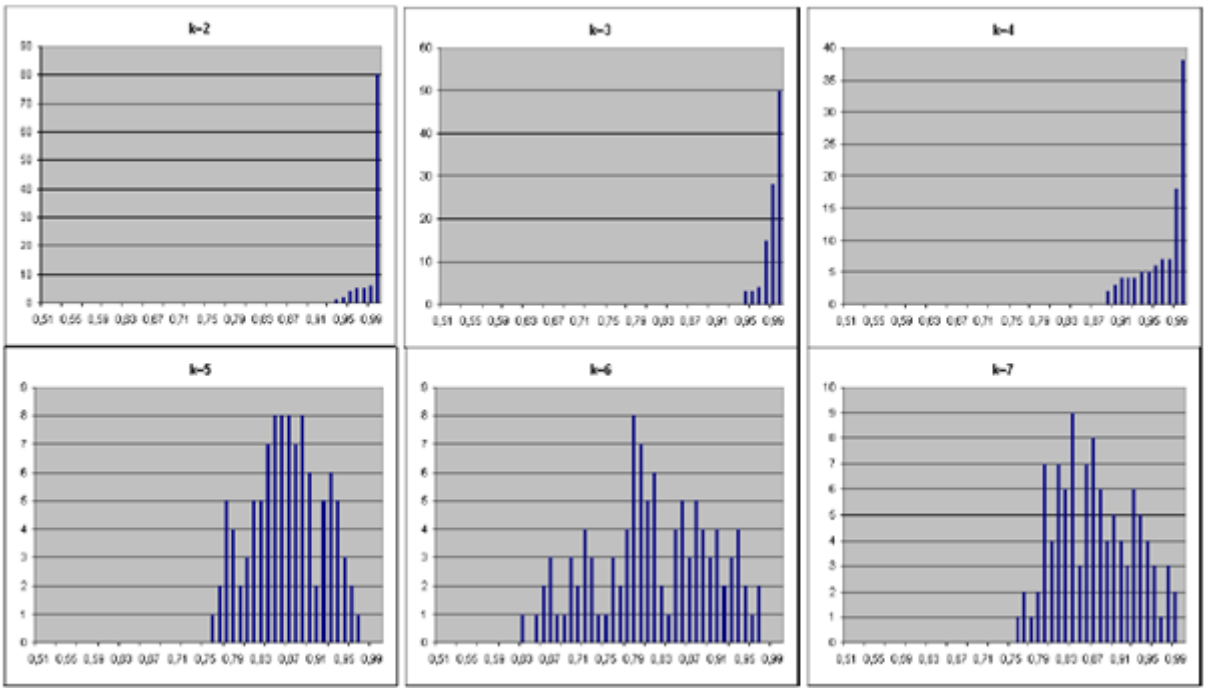


Figura 1: Valores silhouette para conjunto de dados aleatórios (VALENTE, 2013)

5 Critérios e métricas de avaliação

Assim como (DUARTE, 2008) explica, nenhum algoritmo de agrupamento possui um bom desempenho em todos os conjuntos de dados, a sinergia entre os resultados de vários algoritmos pode ser utilizada para que as qualidades dos diferentes algoritmos compensem suas fraquezas.

A análise das características presentes em um conjunto de dados permite a descoberta de padrões que ajudem a compreender o processo que gerou os dados. Muitas dessas características podem ser obtidas por meio da aplicação de fórmulas estatísticas simples,(FRANÇA, 2014) ou então serem observadas por meios de técnicas de visualização, como gráficos, imagens computacionais, modelagens e figuras, etc.

Conhecer os dados com os quais trabalhamos é um fator que auxilia na escolha dos métodos utilizados para a avaliação. Os dados podem ser classificados em quantitativos e qualitativos. Dados quantitativos são representados por valores numéricos. São discretos e contínuos. Os dados qualitativos contem valores nominais e categóricos.(DEMCHENKO; LAAT; MEMBREY, 2014)

Podemos usar métricas de avaliação probabilísticas e uma análise que permita uma visão ampla e detalhada de diversos cenários para cada um dos algoritmos citados na seção quatro, e usando técnicas de inteligência computacional e modelos probabilísticos podemos classificá-los quanto a qualidade dos agrupamentos gerados, tempo de resposta, manipulação de dados, etc. Com estes resultados podemos estudar e aplicar técnicas que nos permitam obter respostas mais rápidas e um desempenho melhor.

Portanto, para avaliar se uma classificação é apropriada, pode-se comparar a variância que deverá ser pequena em relação a divisão. Caso adequada com a variância que deverá ser grande, se a classificação em categorias for boa, isto significa que uma boa divisão do conjunto dos dados em grupos ou categorias é aquela onde os elementos de uma mesma categoria são os mais parecidos entre si e os elementos dos grupos concorrentes são o mais diferente possível. A robustez dos grupos de objetos usados deve ser verificada e, para tanto, é possível realizar uma análise de variância de padrão intergrupos e pelo coeficiente de discriminação.(ARORA; CHANA, 2014)

5.1 Matrizes de afinidade

Para um conjunto de dados constituído por N amostras x , os elementos sij da matriz de Afinidade S = $[s_{ik}]$ contêm a medida da afinidade dos pares de padrões (x_i, x_j), uma vez que a afinidade é definida como sendo a semelhança baseada na relação ou

conexão eventual. (DUARTE, 2008) Devido a propriedades reflexivas, S é simétrica, e com isso $s_{ij} = s_{ji}$. Em geral, nos problemas de agrupamento, S pode ser representada como uma matriz simétrica, e a similaridade é normalmente dada pela distância da métrica reflexiva.

Evidentemente as submatrizes presentes na diagonal principal são bem definidas, possuindo maior magnitude, do que a matriz fora da diagonal principal. Isto é devido a representatividade das afinidades internas e compactação dos dados, portanto a atribuição de agrupamentos ótima se torna aquela que maximiza as magnitudes da diagonal principal e minimiza as demais. (VALENTE, 2013)

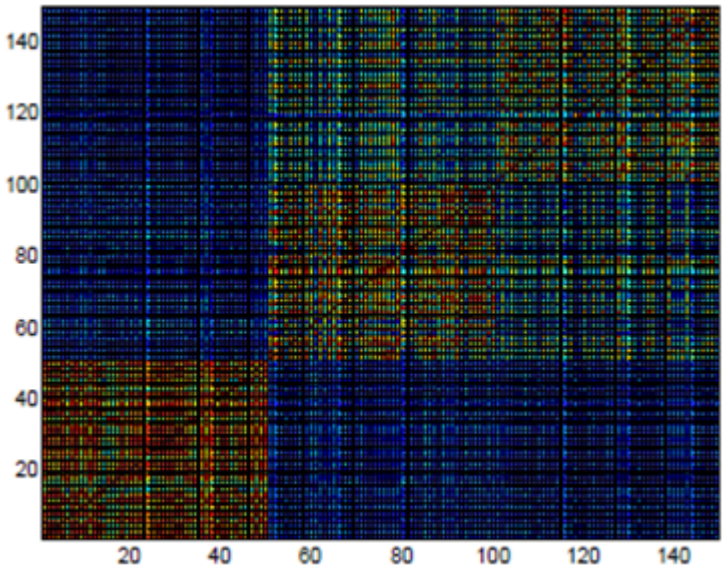


Figura 2: Exemplo de matriz de afinidade (RAMOS, 2014)

Parte II

Resultados

6 Testes e Resultados

Os testes foram feitos em uma máquina *Intel Core* i3 x64 2.27GHZ, com 4GB de memória RAM e capacidade de disco de 500 GB, com sistema operacional *Windows 7*.

Foram utilizados algoritmos em MatLab, visando maximizar e simplificar o trabalho com matrizes. Efetuando agrupamentos com valores que variavam entre dois e dez grupos, visando descobrir o valor de k que mais se adequaria aos dados. Os algoritmos foram implementados seguindo os passos dos pseudocódigos presentes na seção três do presente trabalho.

Utilizando as bases de dados HHAR, TV News Channel e YMVG pudemos simular grandes bases de dados, que se encaixam nos padrões e perfis descritos na seção um, sobre grandes bases de dados. Porém tivemos que fazer algumas adaptações nas bases para que pudéssemos trabalhar em cima dos dados. Como trabalhamos com dados quantitativos e alguns valores eram qualitativos, tivemos que realizar essa transformação nas bases, mas preservando seus valores e pesos para as bases como um todo. Além disso, algumas colunas e linhas tiveram que ser cortadas pois alguns dados estavam faltando, conforme informados pelas Tabelas 1, 2 e 3. Mas esses cortes não influenciaram em perda de informação significativa para as bases de dados, que continuaram dentro dos padrões de grandes bases de dados.

Esses cortes foram feitos por meio de editores de texto, onde foi verificado dados faltosos dessas colunas. O meio mais fácil para ser feito era verificar em cada linha, quantas colunas a menos esta possuia das demais. Assim que era idetificada esta linha então a mesma era cortada da base de dados. Este foi um passo prévio tomado para ser trabalhado nos algoritmos.

Os algoritmos foram escritos para que as entradas fossem do mesmo estilo das bases de dados. Ou seja, em uma tabulação separada por vírgulas. Onde cada separação condizia a uma coluna, ou atributo diferente de uma linha, ou instância.

Ao fazermos o agrupamento nos algoritmos, também adicionamos o método de validação silhouette. Com isso cada resultado que era obtido, as saídas que eram os dados, *labels*, eram verificadas com o método. Já demonstrando a confiabilidade das saídas geradas.

Para garantirmos a veracidade e confiabilidade dos dados, verificando assim o nível dos grupos criados pelas técnicas de agrupamento. Como descrito pelo método, temos valores para cada conjunto de dado, comparando-o com seus semelhantes dentro do mesmo grupo, e com dados dos diferentes grupos. Nos dando resultados que variam entre um e menos um. Para cada valor de k construímos uma média com todos os valores de silhouette

mostrados, retornando o valor dessa média como sendo o resultado final do método de validação. Com este valor fomos capazes de quantificar a qualidade dos agrupamentos criados.

Outro valor relevante para nossa avaliação foi o desvio padrão entre os atributos dos agrupamentos, assim pudemos constatar o nível de covariância entre os mesmos. Esse desvio padrão foi calculado fazendo uma média da distância entre os valores presentes dentro de um mesmo agrupamento.

Por último construímos graficamente a matriz de afinidade, para conseguirmos visualizar os agrupamentos de maneira a entender suas respostas retornadas. Como explicado na seção seis, as matrizes cuja diagonal principal possuem maior grau de magnitude e definição, são os resultados ótimos buscados. A matriz de afinidade é contruída pegando os rótulos gerados na saída e relacionando-os através do próprio MatLab como uma coluna dos próprios dados, e com isso utilizando o algoritmo da matriz era possível montar a mesma.

Ao conseguirmos os rótulos, conseguimos relacionarmos os mesmos ao conjunto de dados para montarmos a matriz de afinidade, e assim verificar a qualidade dos grupos criados. Usando os rótulos obtidos para o melhor valor de k, como será demonstrado nas tabelas e figuras abaixo.

Os algoritmos foram repetidos variando a quantidade de grupos que deviam ser formados, foram variados o valor de k entre dois e dez agrupamentos. E cada algoritmo nos deu os seguintes resultados:

6.1 Testes com Base de Dados HHAR

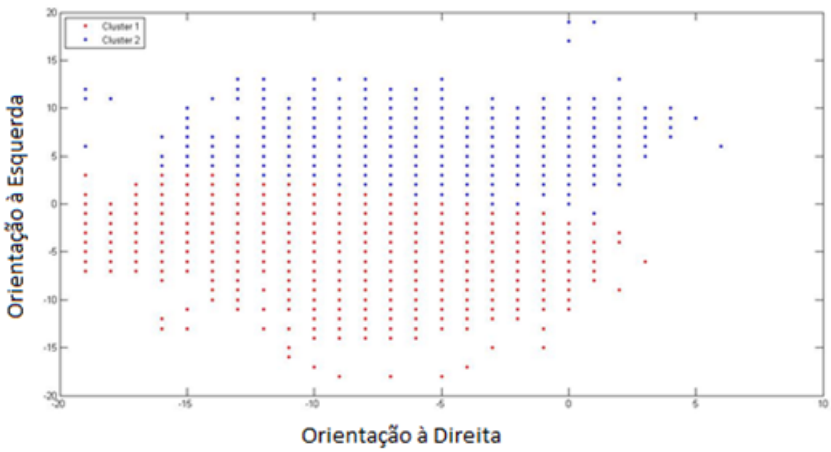
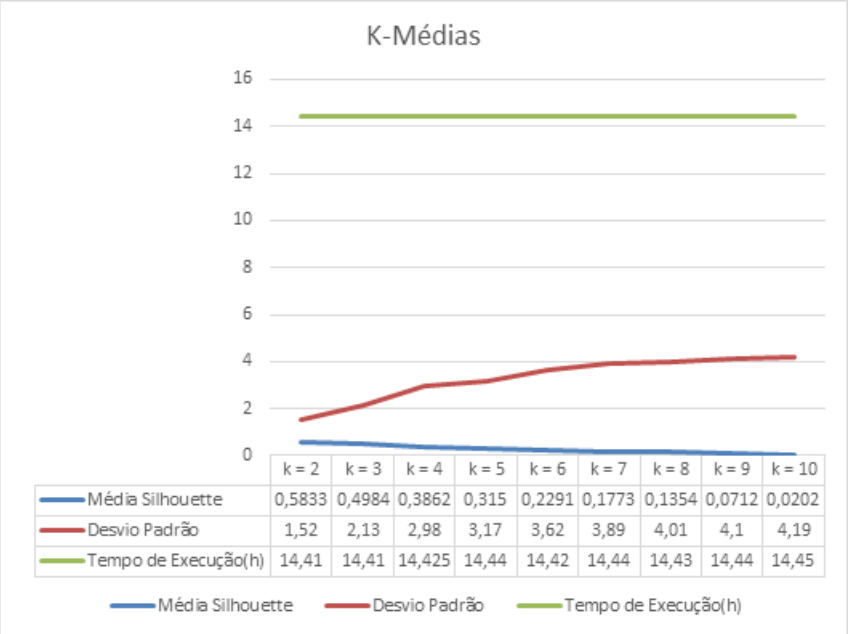


Figura 3: Representação gráfica dos grupos criados pelo algoritmo K-Médias na base de dados HHAR.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Média Silhouette	0,5833	0,4984	0,3862	0,315	0,2291	0,1773	0,1354	0,0712	0,0202
Desvio Padrão	1,52	2,13	2,98	3,17	3,62	3,89	4,01	4,1	4,19
Tempo de Execução(h)	14,41	14,41	14,425	14,44	14,42	14,44	14,43	14,44	14,45

Figura 4: Resultados obtidos para o algoritmo K – Médias na base de dados HHAR.

Como vemos na representação gráfica o método de K-Médias criou dois agrupamentos diferentes baseado nos padrões dos dados. Isto também é ilustrado pelos valores da Figura 4, no qual temos os valores de desvio padrão e uma média dentre os valores dados pelo método de validação silhouette. Notamos que a medida que crescemos a quantidade de grupos, os valores do silhouette diminuíam, demonstrando como os grupos perdiam similaridade, o que podemos dizer, era uma perda de qualidade dos agrupamentos. Ao mesmo tempo também tivemos um aumento do valor de desvio padrão, o que nos mostra como os dados estavam ficando mais distantes uns dos outros à medida que o número de grupos crescia. Por fim outro ponto que notamos foi o do alto custo computacional gasto pelo algoritmo, representado pelo tempo de execução do algoritmo. Mesmo para o agrupamento ótimo para este algoritmo, fora gasto 14,41 horas para a execução do mesmo.

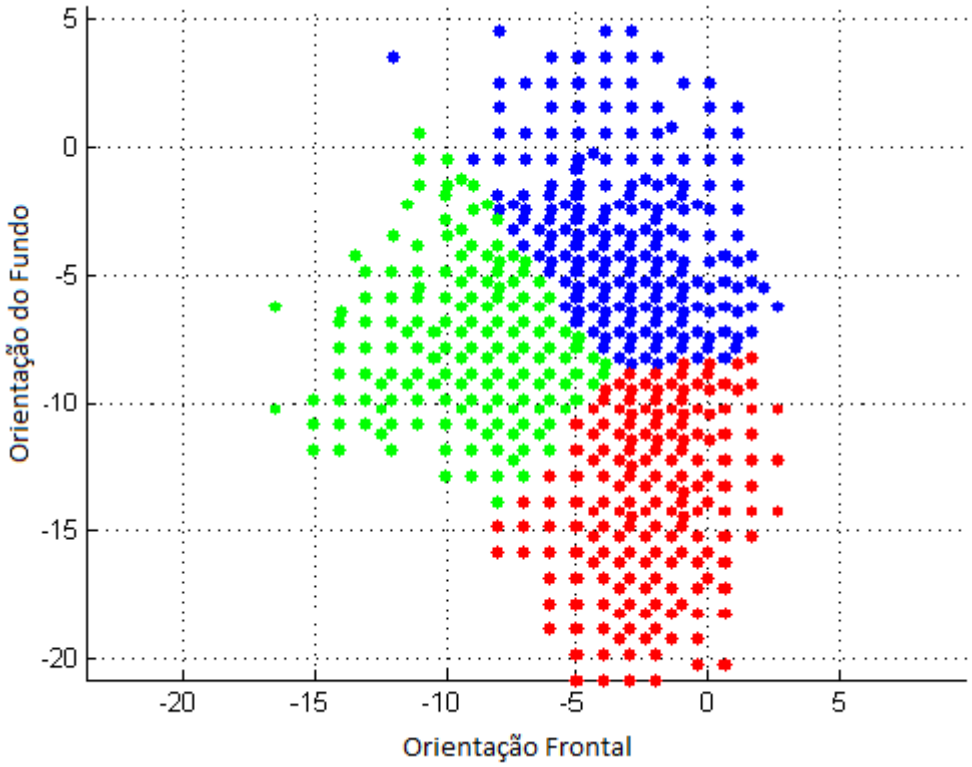
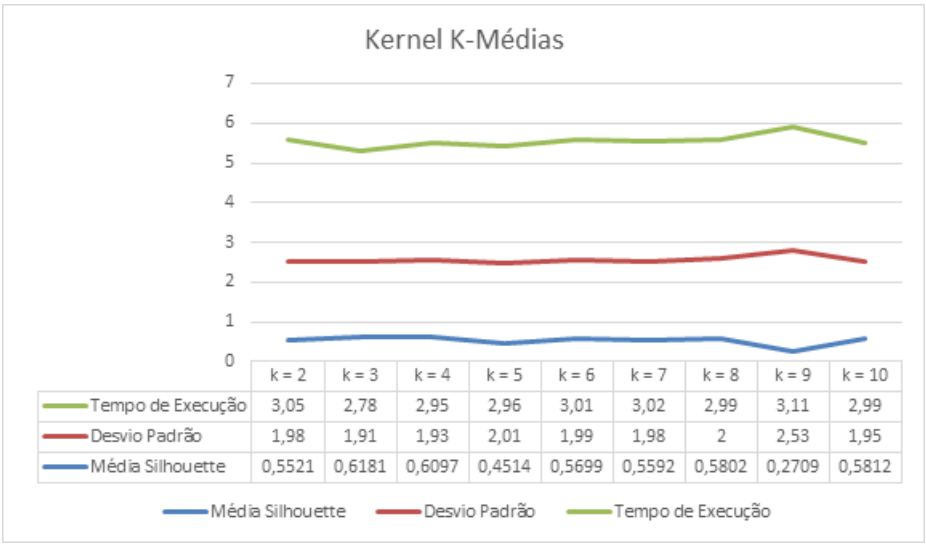


Figura 5: Representação gráfica dos grupos criados pelo algoritmo Kernel K-médias na base de dados HHAR

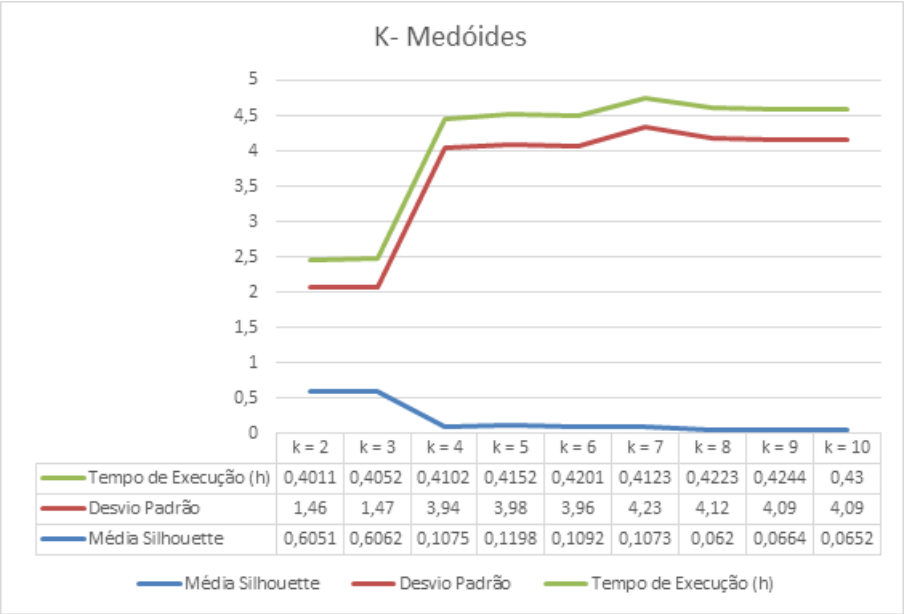


	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	3,05	2,78	2,95	2,96	3,01	3,02	2,99	3,11	2,99
Desvio Padrão	1,98	1,91	1,93	2,01	1,99	1,98	2	2,53	1,95
Média Silhouette	0,5521	0,6181	0,6097	0,4514	0,5699	0,5592	0,5802	0,2709	0,5812

Figura 6: Resultados obtidos para algoritmo Kernel k-médias na base de dados HHAR.

O algoritmo Kernel k-médias se mostrou o que manteve maior média durante todas as iterações de valor de k. A média alta mostra quão eficaz para esta base de dados é este algoritmo. Isso também demonstra a qualidade dos grupos gerados, mesmo para outros valores fora do comum para o caso. Apesar de não possuir a maior média de silhouette, podemos dizer que é o melhor resultado devido à combinação deste, com seu baixo desvio

padrão e tempo regular de execução. A representação gráfica para o valor de melhor média silhouette demonstra como os grupos ficam bem separados, apesar de haver algumas dissimilaridades e erros.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução (h)	0,4011	0,4052	0,4102	0,4152	0,4201	0,4123	0,4223	0,4244	0,43
Desvio Padrão	1,46	1,47	3,94	3,98	3,96	4,23	4,12	4,09	4,09
Média Silhouette	0,6051	0,6062	0,1075	0,1198	0,1092	0,1073	0,062	0,0664	0,0652

Figura 7: Resultados obtidos para o algoritmo K-Medóides na base de dados HHAR.

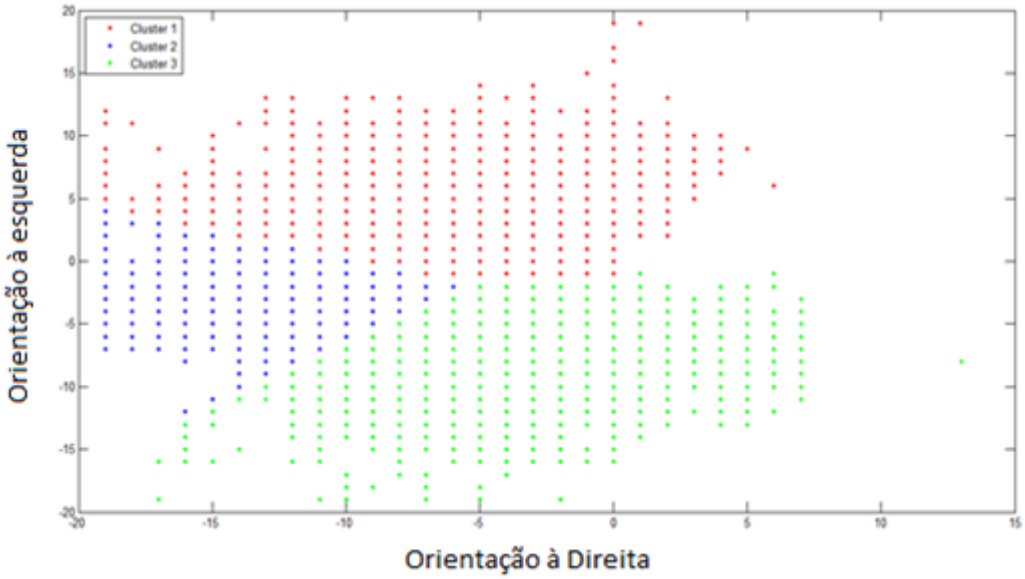


Figura 8: Representação gráfica dos grupos criados para algoritmo K-Medóides na base de dados HHAR.

No algoritmo K-Medóides, notamos que encontramos um valor de agrupamento ótimo diferente do que foi encontrado pelo algoritmo K-Médias. Neste caso, o valor de grupos sobe para três, ao invés de somente dois. Mas analisando bem os dados, vemos que a diferença para o número de agrupamentos semelhante ao algoritmo anterior é muito pequena. No caso do desvio padrão, temos que, para três grupos um valor de 1,46, enquanto para dois grupos temos 1,47. A diferença está nos centésimos. O mesmo acontece para o valor médio do silhouette. Ambos os grupos possuem bom nível de similaridade, com uma diferença de apenas 0,0013 entre um e outro. O algoritmo K-Medóides se mostra mais eficiente que o K-Médias, pois os grupos têm um grau de similaridade maior, menor desvio padrão, e mais importante, um tempo computacional bem abaixo do visto anteriormente. Diferentemente do K-Médias, este algoritmo leva apenas 0,4052 para identificar o número de grupos ótimos para esta base de dados. Porém um ponto negativo que chamou a atenção, foi a perda de desempenho que este algoritmo apresentou para um número K superior a quatro, tanto para o silhouette, quanto para o desvio padrão. Porém o tempo de execução não é afetado do mesmo modo.

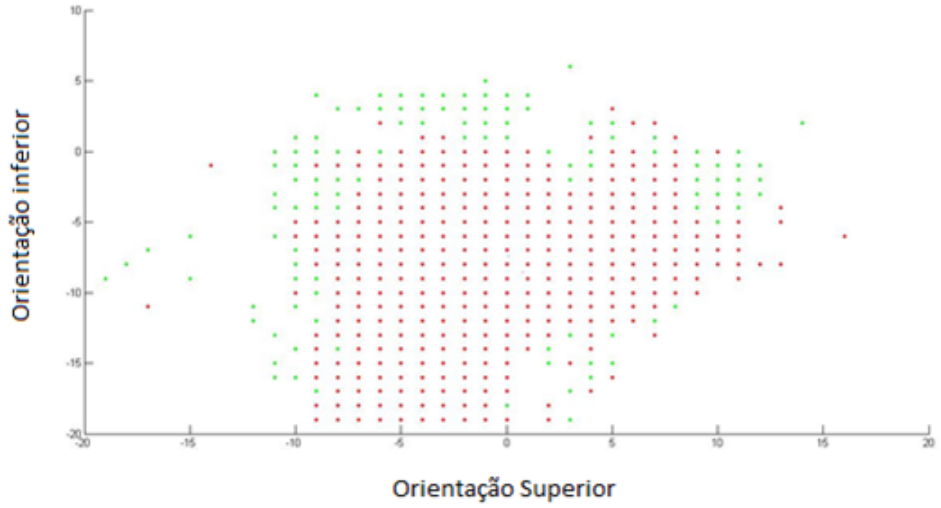
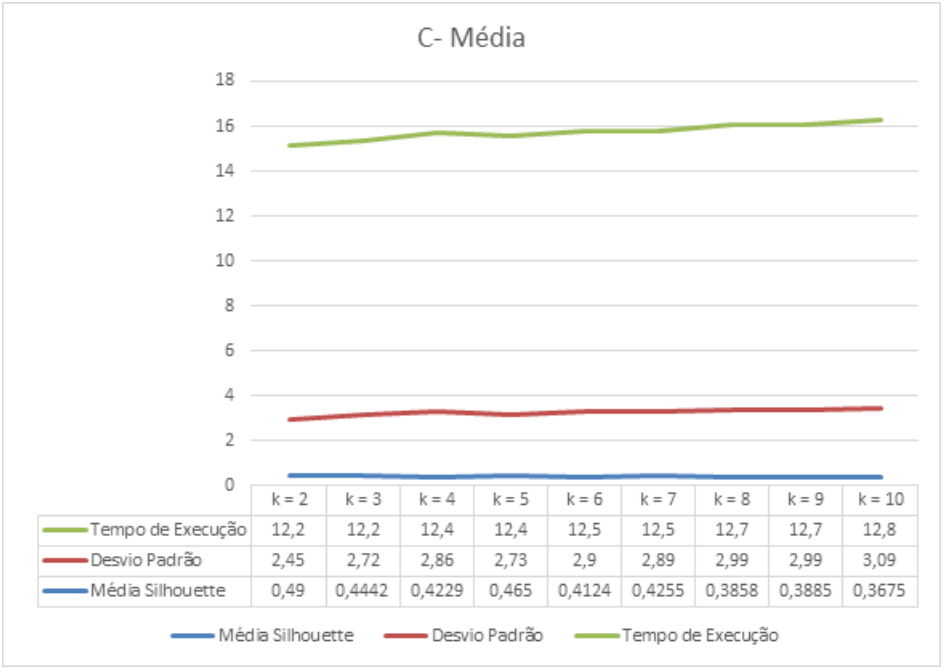


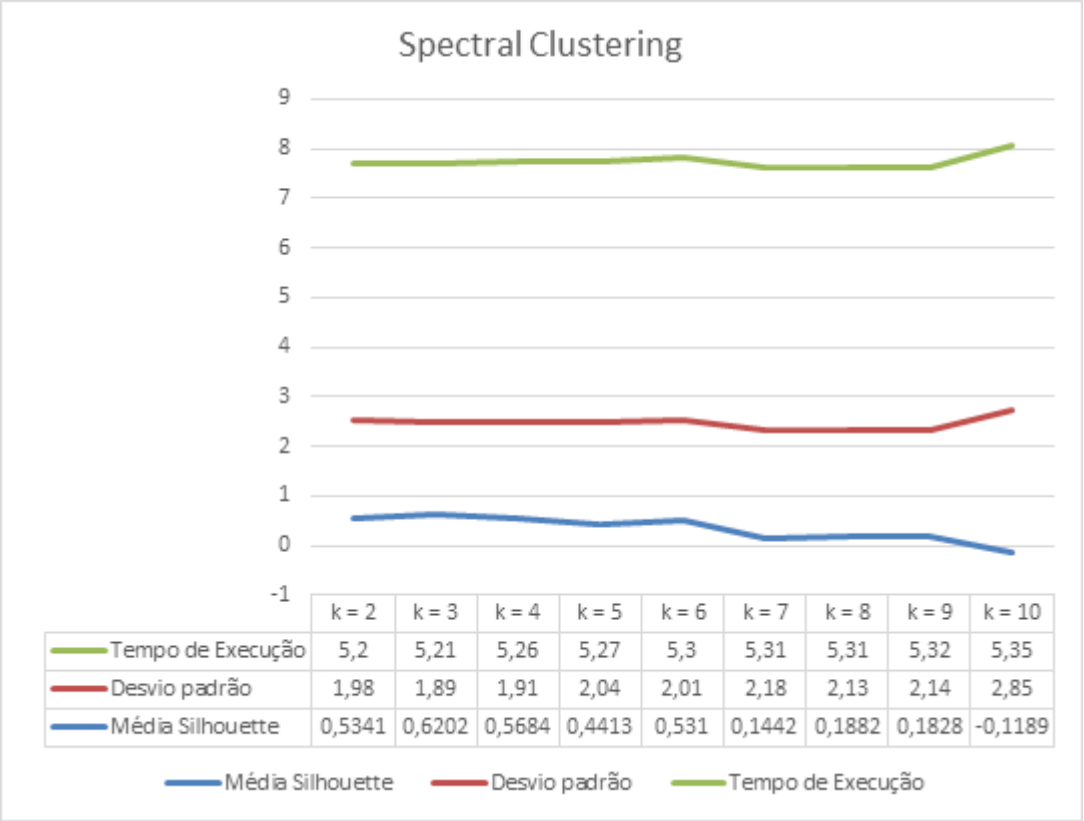
Figura 9: Representação Gráfica dos grupos criados para algoritmo C-Médias na base de dados HHAR.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	12,2	12,2	12,4	12,4	12,5	12,5	12,7	12,7	12,8
Desvio Padrão	2,45	2,72	2,86	2,73	2,9	2,89	2,99	2,99	3,09
Média Silhouette	0,49	0,4442	0,4229	0,465	0,4124	0,4255	0,3858	0,3885	0,3675

Figura 10: Resultados obtidos para algoritmo Fuzzy C-Médias na base de dados HHAR.

No algoritmo C-Médias obtivemos valores aceitáveis para o método de validação, com um número maior para um agrupamento de somente dois padrões. O desvio padrão menor também ocorreu para este valor, sendo assim temos um agrupamento ótimo neste caso para dois grupos apenas. O tempo de execução para este algoritmo foi quase tão grande quanto o K-Médias, chegando a uma média de 12,2 horas para chegarmos a um resultado. Porém esta medida tem pouca variação à medida que se varia o número de agrupamentos. Portanto podemos dizer que este número em si não influencia nessa variável.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	5,2	5,21	5,26	5,27	5,3	5,31	5,31	5,32	5,35
Desvio padrão	1,98	1,89	1,91	2,04	2,01	2,18	2,13	2,14	2,85
Média Silhouette	0,5341	0,6202	0,5684	0,4413	0,531	0,1442	0,1882	0,1828	-0,1189

Figura 11: Resultados obtidos para o algoritmo Spectral Clustering na base de dados HHAR.

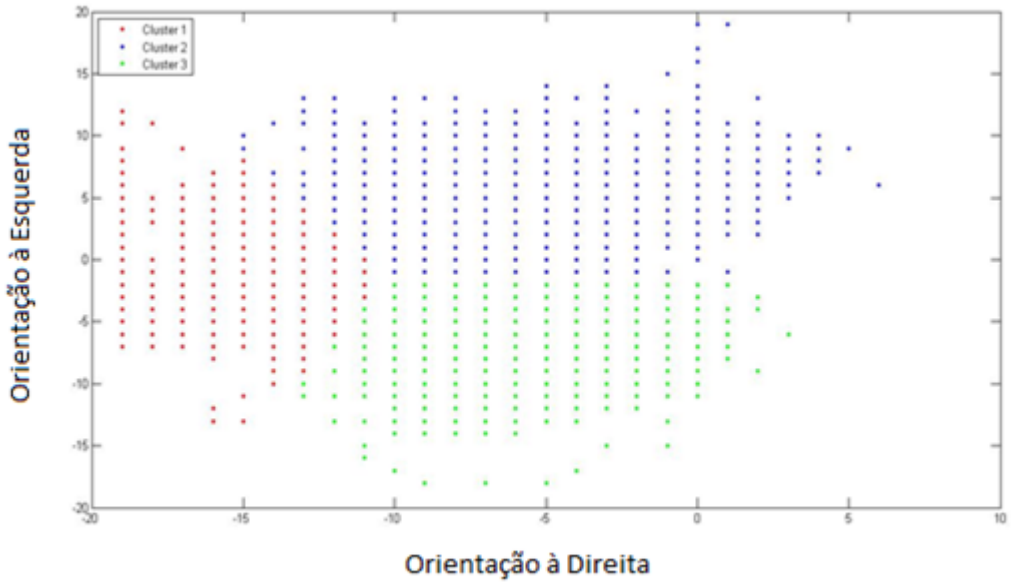


Figura 12: Representação gráfica dos grupos criados para algoritmo Spectral Clustering na base de dados HHAR.

O algoritmo *Spectral Clustering* apresentou bons resultados, também para um número de agrupamentos igual a três. Este método foi bastante eficaz, tendo obtido o

maior grau de validação dentre os algoritmos utilizados. Apesar do valor de desvio padrão não ser o menor, os grupos ficaram bem separados possuindo alta simetria dentre seus membros, e alta dissimilaridade com os demais. O valor de desvio padrão maior pode ser explicado devido à grande quantidade de valores presentes na base de dados. O tempo computacional foi razoável, visto que o K-Médias teve um valor três vezes maior, mas também tivemos algoritmos conseguindo executar em um tempo menor. Notou-se também que o tempo não apresentou grande variação, mesmo com diferentes valores de K. E um aspecto importante foi que este algoritmo foi o único a apresentar um valor de silhouette negativo, mostrando que os agrupamentos formados estavam sendo entendidos, e que realmente os padrões existentes não suportavam grandes quantidades de grupos.

Para efeitos de comparação podemos montar uma tabela com os valores de tempo de execução de cada algoritmo para a base de dados HHAR, para os melhores agrupamentos obtidos. O número k que simboliza o melhor conjunto de agrupamentos é aquele cuja média silhouette apresenta maior valor. Sendo assim temos:

Tabela 5: Valores de tempo para melhores agrupamentos na base HHAR

Algoritmo	**Tempo de Execução**
K-Médias	14,41 horas
K-Medóides	0,40 horas
Kernel K-Médias	3,18 horas
Fuzzy C-Médias	12,2 horas
Spectral Clustering	5,21 horas

No quesito tempo de execução vemos que o K-Medóides consegue superar amplamente os demais algoritmos com um tempo de execução de aproximadamente 87 porcento mais rápido que o kernel k-médias que tem o segundo melhor tempo. O Spectral Clustering, dito na literatura, ter um bom desempenho não expressou isso pelo tempo de execução apesar de ter alcançado a melhor média silhouette.

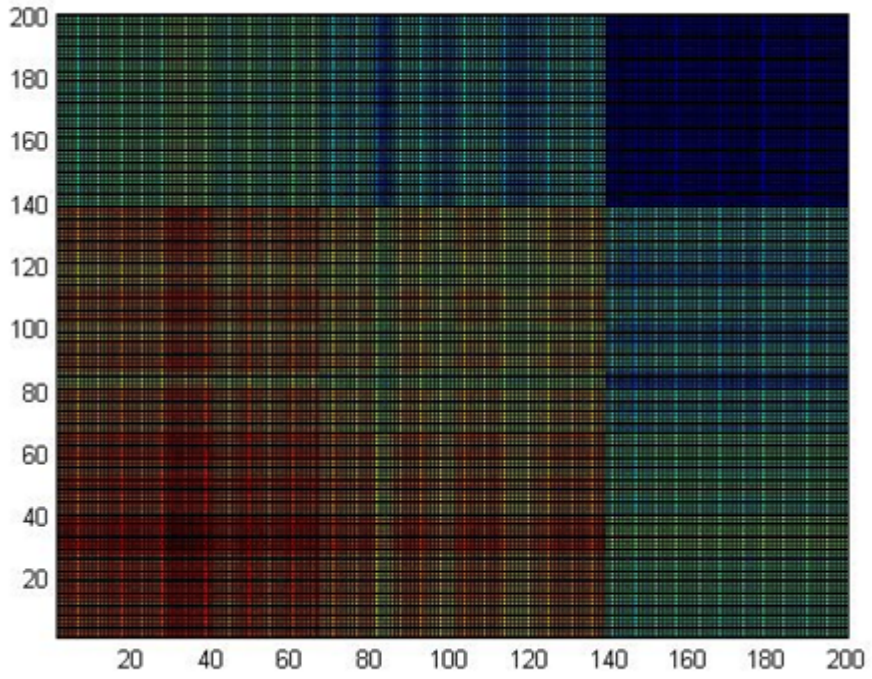


Figura 13: Matriz de Afinidade para agrupamentos formados com valor de k = 2 com algoritmo K-Médias na base de dados HHAR.

Vemos que para a matriz de afinidade do K-Médias, temos uma diagonal principal relativamente correlacionada, onde valores mais altos tem uma dissociação menor. Os primeiros grupos são menos similares que os terminais. E os resultados aparentaram ser satisfatórios, visto que, a base de dados tinha uma considerável quantidade de dados.

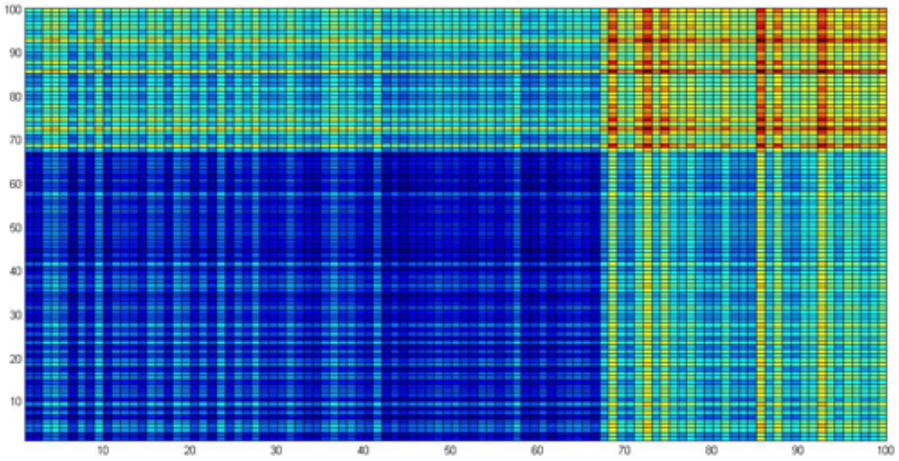


Figura 14: Matriz de Afinidade para agrupamentos formados com valor de k = 2 com algoritmo Kernel K-Médias na base de dados HHAR.

Os grupos formados com o algoritmo Kernel K-Médias se mostram bem similares,

visto que temos na matriz de afinidade construída para dois agrupamentos uma diagonal principal com maioria de valores coesos. Principalmente os primeiros grupos onde estamos vendo que os dados estão bem compactos. Porém algum dado fora dessa diagonal apresentam um comportamento semelhante, o que tira um pouco do alto grau para os grupos diferentes.

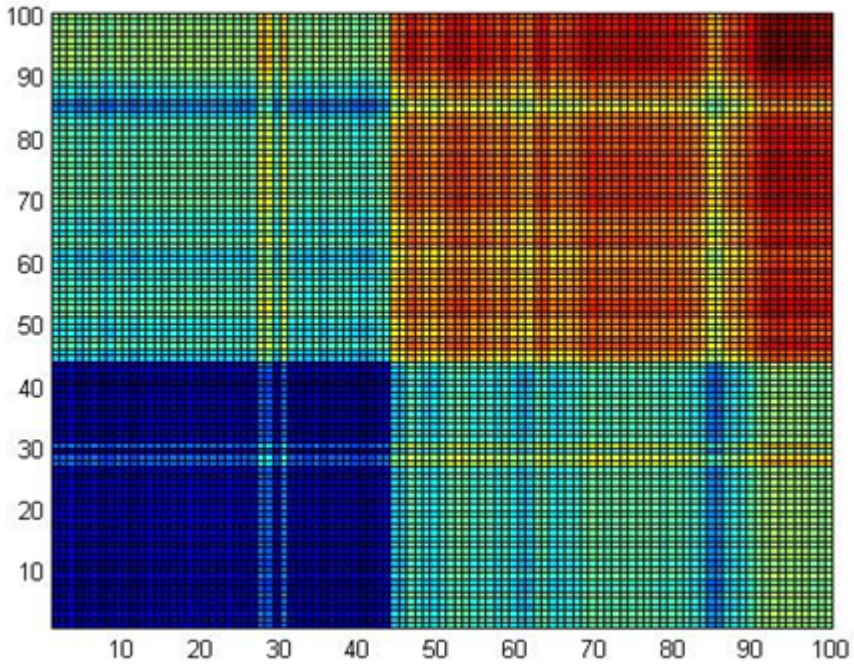


Figura 15: Matriz de Afinidade para agrupamentos formados com valor de k = 2 com algoritmo K-Medóides na base de dados HHAR.

Para a matriz obtida para o algoritmo K-Medóides temos que os grupos finais possuem menos similaridades do que os primeiros, contrário ao algoritmo K-Médias. Mas como foi notado, esta matriz está mais nítida do que a anterior, denotando um grau de coesão menor. Os grupos estão mais separados, apesar de notarmos alguma similaridade fora dos grupos de origem, o que demonstra algum erro, mas ainda assim os resultados também podem ser considerados não satisfatórios.

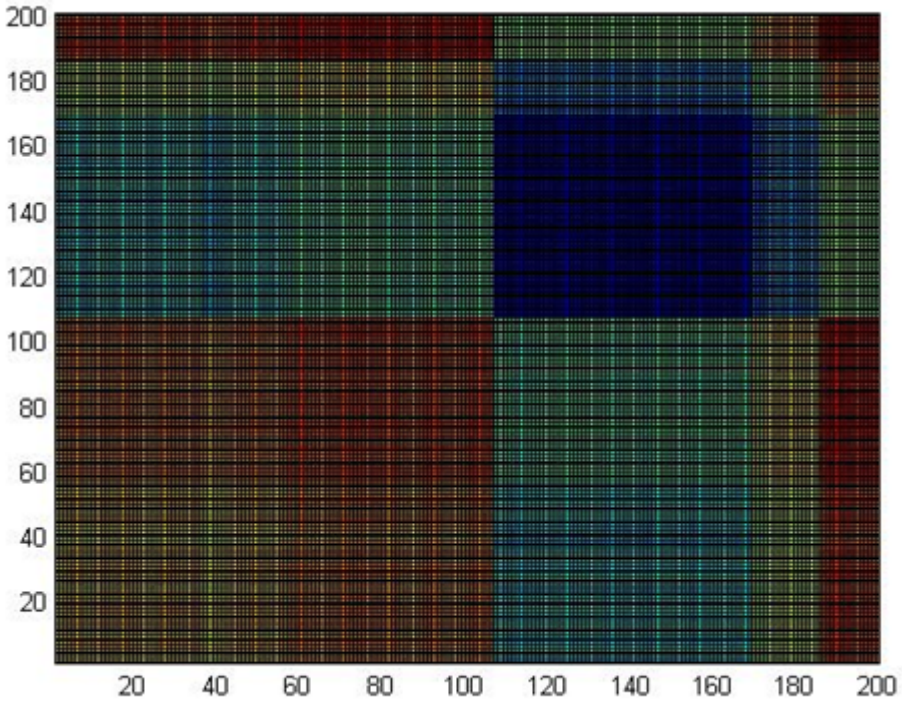


Figura 16: Matriz de Afinidade para agrupamentos formados com valor de k = 3 com algoritmo Fuzzy C-Médias na base de dados HHAR.

O algoritmo C-Médias apresentou uma matriz, que deixou transparecer ser um meio termo entre os dois algoritmos anteriores, K-Médias e K-Medóides. Tendo a diagonal principal e alguns pontos fora da mesma, valores não muito bem correlacionados, sobressaindo os primeiros grupos. Os pontos que foram observados, relacionando a matriz com os resultados obtidos, comprovam que este algoritmo não foi tão eficiente quanto os outros, mesmo podendo considerar resultados medianos para seu uso. A dissociação notada fora da diagonal principal não pode deixar de ser considerada.

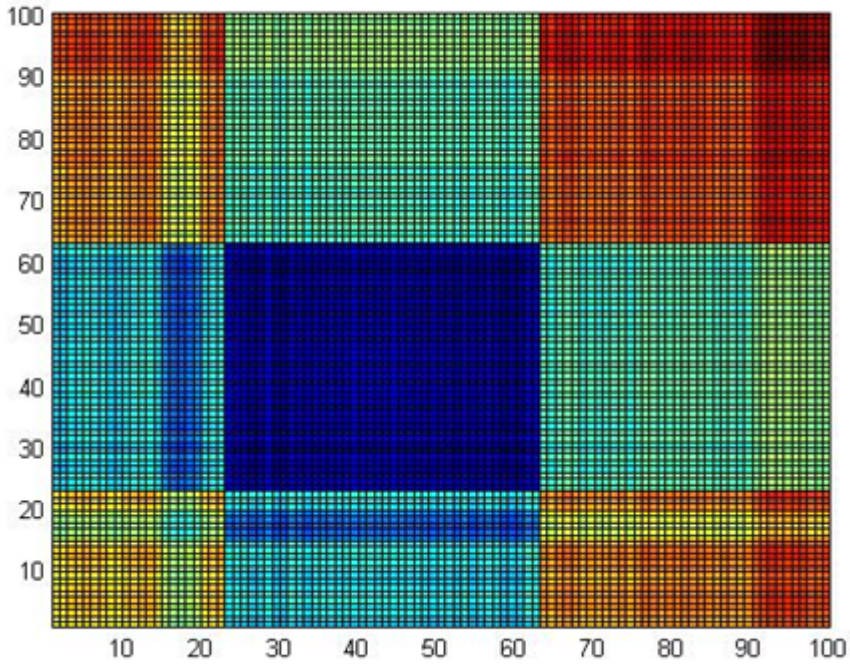


Figura 17: Matriz de Afinidade para agrupamentos formados com valor de k = 5 com algoritmo Spectral Clustering na base de dados HHAR.

A matriz para o algoritmo de *Spectral Clustering* foi a que apresentou maiores graus de correlação fora da diagonal principal. Porém também foi aquela em que esses graus foram maiores tanto nos primeiros quanto para os últimos grupos. Tendo valores medianos um valor dissociativo. Mas, pelos valores quantitativos recebidos, este algoritmo apresentou bom desempenho e resposta aceitável para esta base de dados, mostrando ser um bom método, assim como a literatura e revisões teóricas refutavam.

6.2 Testes com Base de Dados TV News Channel Detection

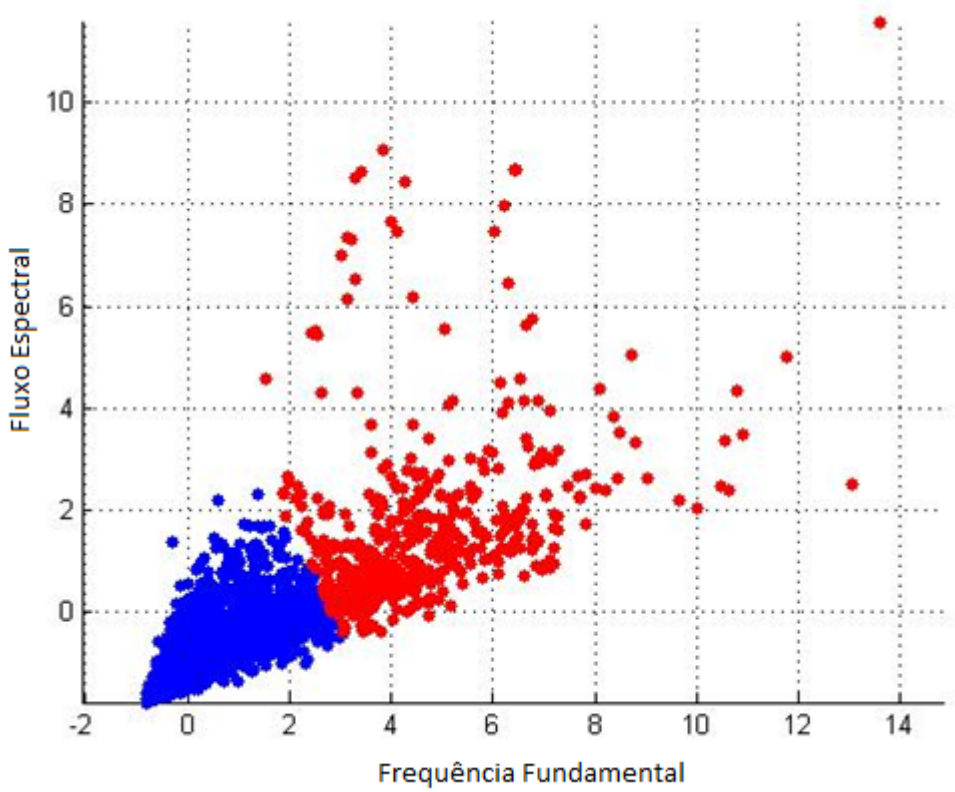
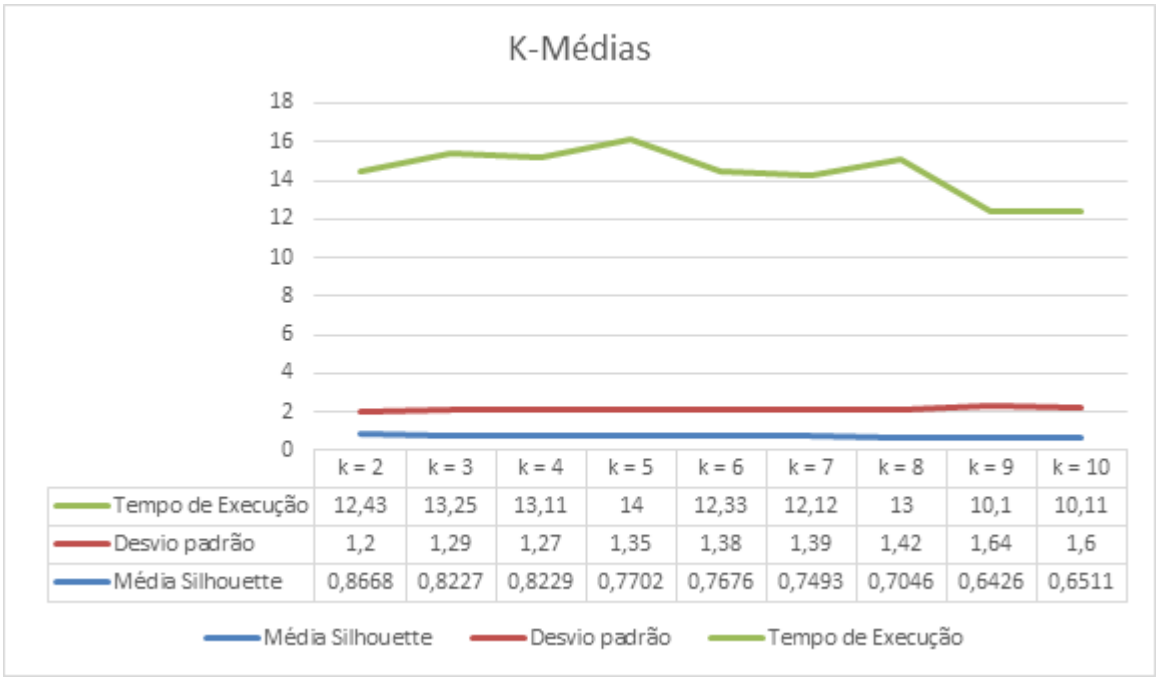


Figura 18: Representação gráfica dos grupos formados com algoritmo K-Médias na base de dados TV News Commercial Detection.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	12,43	13,25	13,11	14	12,33	12,12	13	10,1	10,11
Desvio padrão	1,2	1,29	1,27	1,35	1,38	1,39	1,42	1,64	1,6
Média Silhouette	0,8668	0,8227	0,8229	0,7702	0,7676	0,7493	0,7046	0,6426	0,6511

Figura 19: Resultados obtidos para algoritmo K-Médias com base de dados TV News Commercial Detection.

Com a base de dados TV News utilizando o algoritmo K-Médias percebemos

uma peculiaridade que só aconteceu nesse caso. A qualidade dos agrupamentos para diferentes valores de k ficaram extremamente constantes, sem mencionar que foram ótimos resultados considerando o tamanho da base. Com isso também temos valores constantes de desvio padrão, o que mostra que os dados tendem a estar em distancias fixas entre os agrupamentos. Apesar desses altos valores, o tempo de execução permaneceu alto, do mesmo modo que aconteceu com a base de dados anterior. O tamanho da base diminui consideravelmente, porém permanece alto, o número de atributos aumenta por outro lado. Algo que faz o tempo em média cair por duas horas, mas ainda assim são tempos considerados altos.

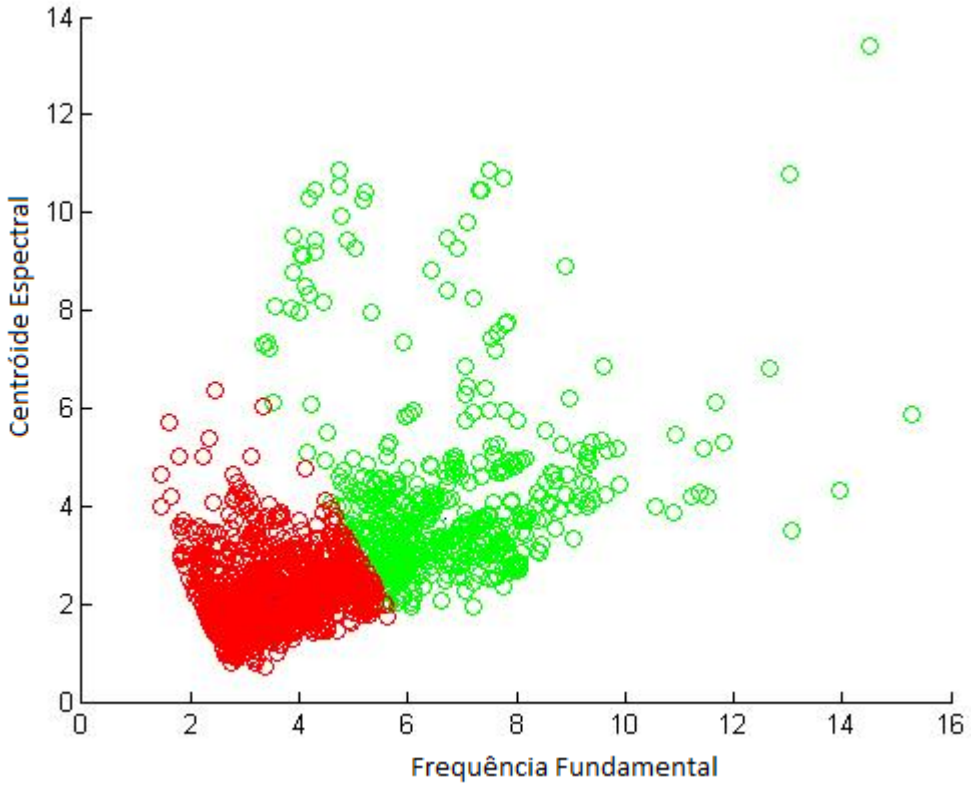
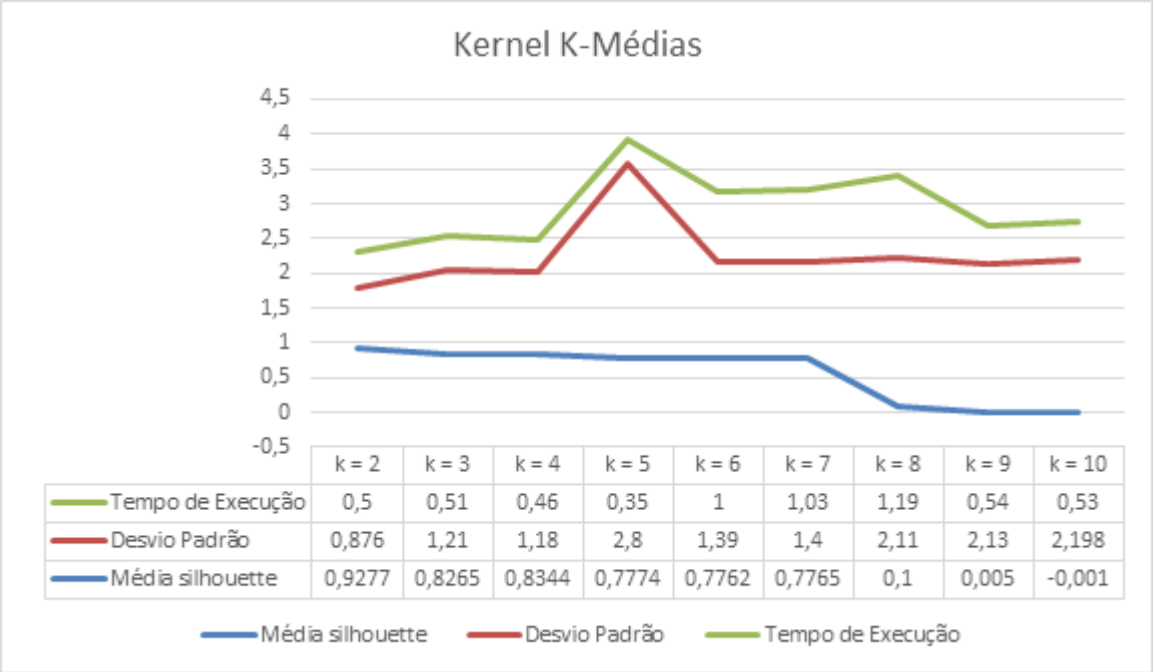


Figura 20: Representação gráfica dos grupos formados com algoritmo Kernel K-Médias na base de dados TV News Commercial Detection.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	0,5	0,51	0,46	0,35	1	1,03	1,19	0,54	0,53
Desvio Padrão	0,876	1,21	1,18	2,8	1,39	1,4	2,11	2,13	2,198
Média silhouette	0,9277	0,8265	0,8344	0,7774	0,7762	0,7765	0,1	0,005	-0,001

Figura 21: Resultados obtidos com algoritmo Kernel K-Médias para base de dados TV News Commercial Detection.

O algoritmo Kernel K-Médias obteve a maior média silhouette dentre todos os algoritmos testados nesta base de dados. Consequentemente também conseguiu o menor desvio padrão. O algoritmo tem uma pequena queda de qualidade a medida que variávamos o valor de k. Porém, quando o número de agrupamentos ficava acima de sete, tínhamos uma queda brusca de desempenho. Os agrupamentos ficavam muito dissemelhantes e a média se tornava negativa. Como podemos ver na Figura 21. O tempo de execução teve bons resultados, descobrindo o melhor valor de agrupamento com cinquenta minutos. Para uma grande base de dados este é um valor aceitável. A queda no desempenho a números maiores de quantidade de grupos não era acompanhada por um aumento de tempo de execução, que nunca ultrapassará mais do que uma hora e vinte minutos.

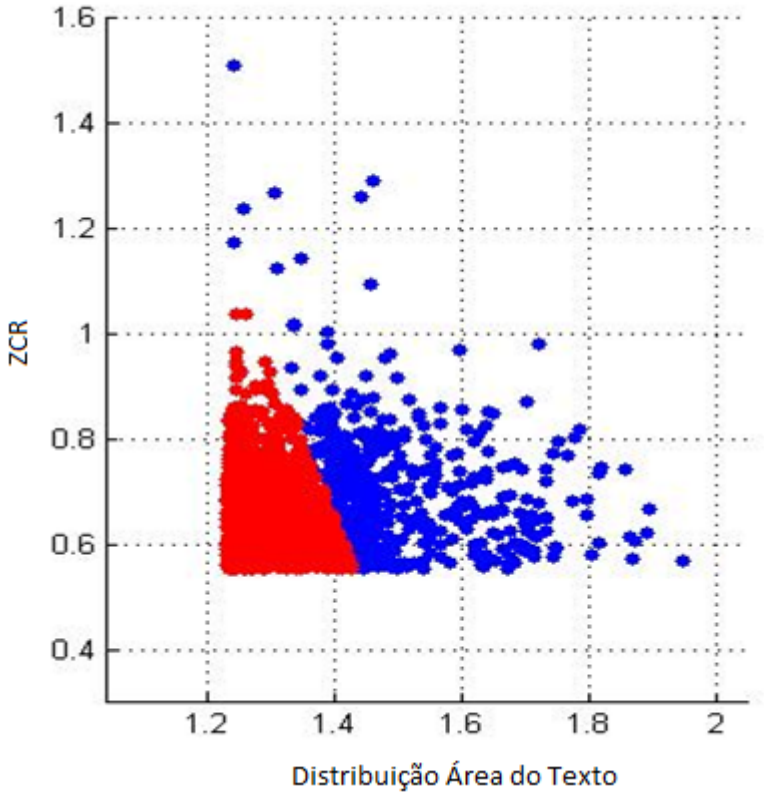
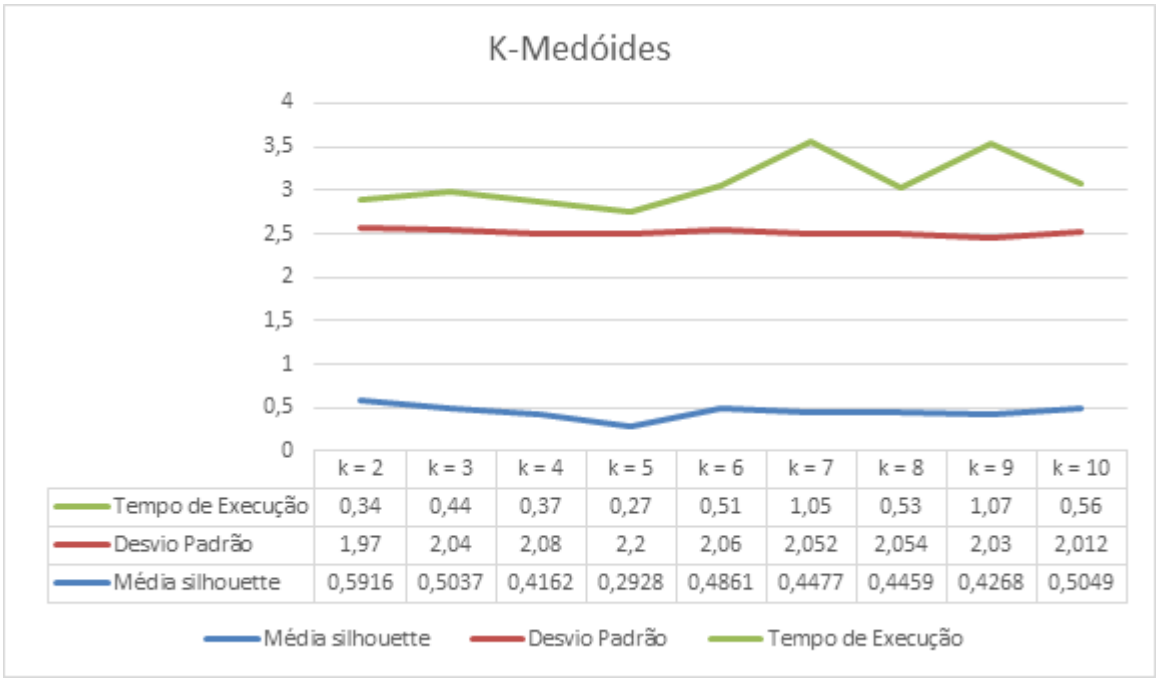


Figura 22: Representação gráfica dos agrupamentos formados com algoritmo K-Medóides na base de dados TV News Commercial Detection.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	0,34	0,44	0,37	0,27	0,51	1,05	0,53	1,07	0,56
Desvio Padrão	1,97	2,04	2,08	2,2	2,06	2,052	2,054	2,03	2,012
Média silhouette	0,5916	0,5037	0,4162	0,2928	0,4861	0,4477	0,4459	0,4268	0,5049

Figura 23: Resultados obtidos para algoritmo K-Medóides para base de dados TV News Commercial Detection.

Na base de dados HHAR o algoritmo K-Medóides obteve bons resultados, em respostas mais rápidas que os demais algoritmos e uma boa média silhouette. Contudo

esse desempenho não se repetiu com a base TV News. Apesar do tempo de execução continuar sendo o menor dentre os algoritmos testados, o valor da média não excedeu 0,6. Apesar de conter mais similaridades entre os grupos do que o contrário, estes foram os piores resultados dentre os demais. O que pode ser considerado de positivo é que o valor permanece numa baixa variação, tendo o menor valor para um número de grupos igual a cinco. O desvio padrão acompanha a variação da média silhouette. Com isso pode-se verificar como o desempenho é totalmente inerente a base de dados e seus atributos. O grau de correlação, ou o modo como são distribuídos os dados devem ser considerados.

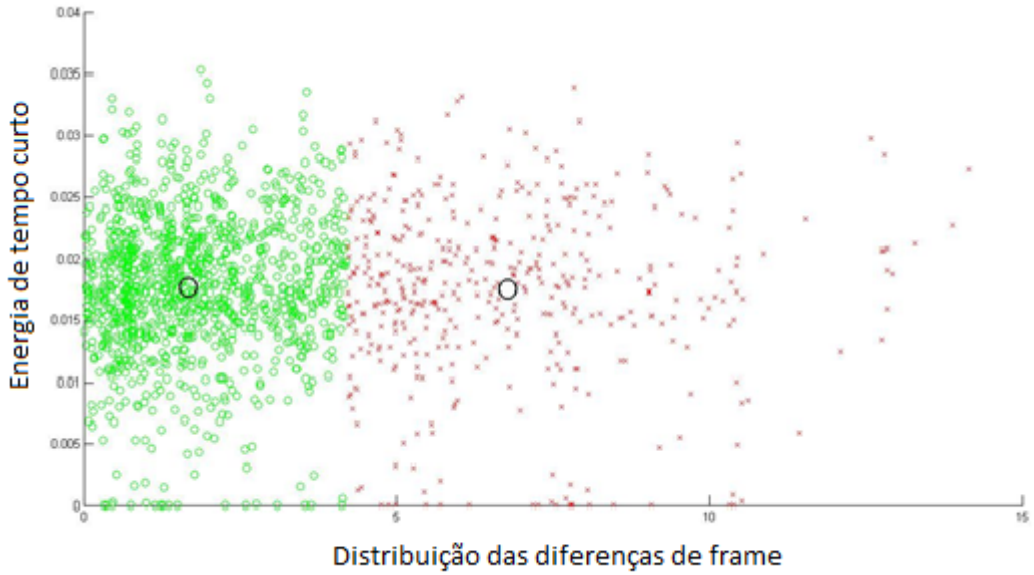
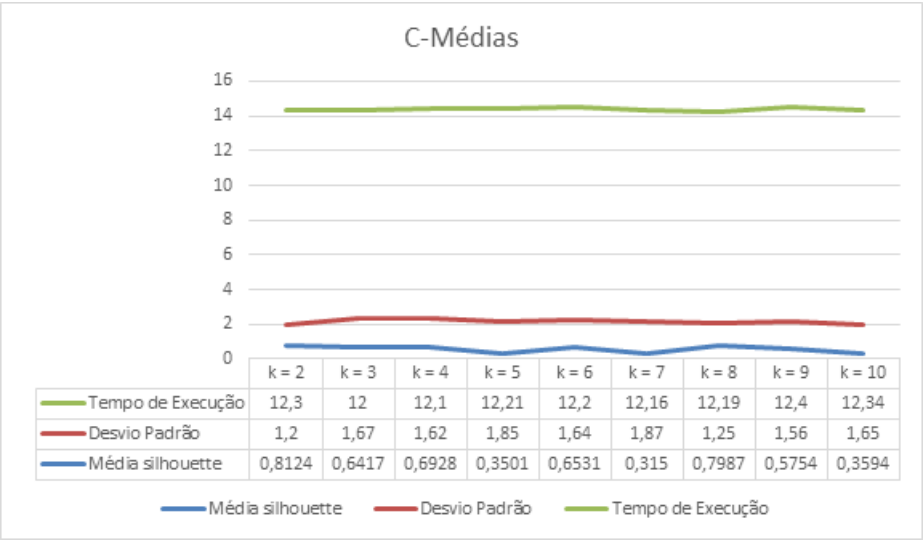


Figura 24: Representação gráfica dos grupos formados com algoritmo Fuzzy C-Médias na base de dados TV News Commercial Detection.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	12,3	12	12,1	12,21	12,2	12,16	12,19	12,4	12,34
Desvio Padrão	1,2	1,67	1,62	1,85	1,64	1,87	1,25	1,56	1,65
Média silhouette	0,8124	0,6417	0,6928	0,3501	0,6531	0,315	0,7987	0,5754	0,3594

Figura 25: Resultados obtidos para algoritmo Fuzzy C-Médias para base de dados TV News Commercial Detection.

O algoritmo Fuzzy C-Médias segue a direção oposta do algoritmo K-Medóides. Os resultados para a base de dados TV News foram bem melhores do que para a base HHAR. Tendo um valor 60 porcento melhor para a média silhouette. Isso também causa uma menor distância entre os dados e consequentemente um menor desvio padrão. Esse número porém só pôde ser dito para o melhor valor de agrupamento encontrado. Para os demais valores obteve-se uma queda de desempenho, exceto para um valor de k igual a oito. Como pode ser observado na Figura 25. Portanto, este algoritmo mostrou grande variação apesar de não conter nenhum valor próximo o suficiente de zero ou negativo. O tempo de execução foi praticamente o mesmo neste caso. Com uma média praticamente igual à dos testes anteriores com este mesmo algoritmo. A variação do tempo é imperceptível, com poucos minutos de diferença entre diferentes valores de k. Mas nestes casos, com bases maiores, estes minutos não indicam grande alarde para o desempenho.

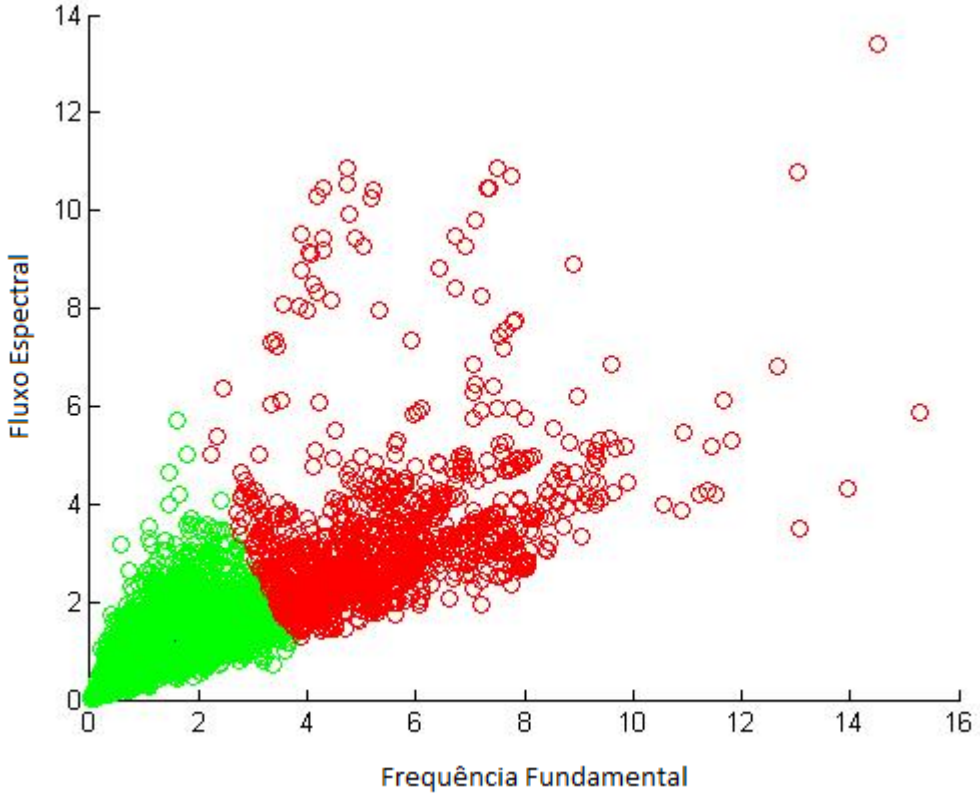
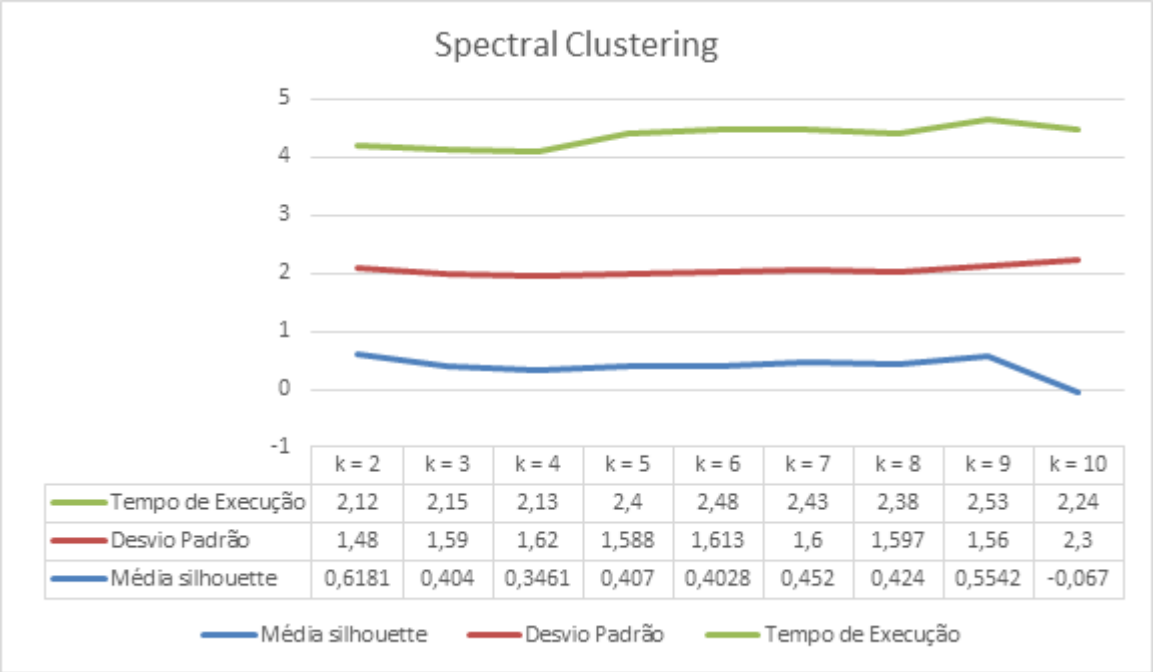


Figura 26: Representação gráfica dos agrupamentos formados pelo algoritmo Spectral Clustering na base de dados TV News Commercial Detection.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	2,12	2,15	2,13	2,4	2,48	2,43	2,38	2,53	2,24
Desvio Padrão	1,48	1,59	1,62	1,588	1,613	1,6	1,597	1,56	2,3
Média silhouette	0,6181	0,404	0,3461	0,407	0,4028	0,452	0,424	0,5542	-0,067

Figura 27: Resultados obtidos com algoritmo Spectral Clustering para base de dados TV News Commercial Detection.

O algoritmo *Spectral Clustering* tem uma média razoavelmente boa para os valores silhouette. No melhor caso, um valor de 0,6181, mostrando uma boa similaridade entre os dados para os agrupamentos com valor de k igual a dois. Mas após isso sua média cai para a casa de 0,40, e sofre pouca variação, com exceção para um número de agrupamentos igual a dez, onde apresenta maior grau de dissimilaridade entre os grupos criados.

Neste caso o algoritmo diminuiu seu tempo de execução em relação a testes anteriores, ficando na casa de duas horas e alguns minutos para nos dar a resposta desejada. Como os dados tem menor quantidade do que a base anterior, e este algoritmo tem na literatura bons resultados, era esperado resultados melhores tanto para o tempo, quanto para a média silhouette.

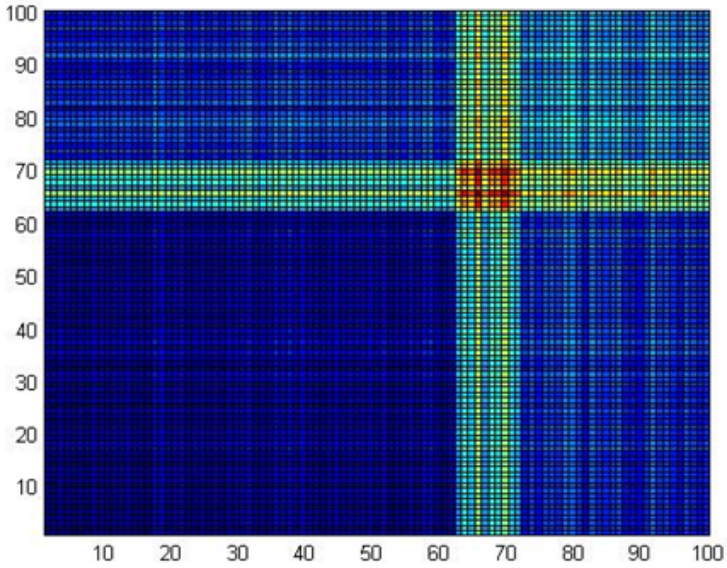


Figura 28: Matriz de Afinidade para agrupamentos formados com valor de k = 2 com algoritmo K-Médias na base de dados TV News Commercial Detection.

Vimos que as matrizes de afinidade para a base de dados TV News estão mais nítidas do que as encontradas para a base de dados HHAR. Com agrupamentos mais definidos e similares entre si, e com menor grau de coesão para agrupamentos diferentes. No primeiro caso, para o algoritmo K-Médias temos uma matriz que tem praticamente toda a diagonal principal num ótimo grau de coesão. Demonstrando a qualidade do grupo e a pouca variação que o algoritmo demonstrou para a base.

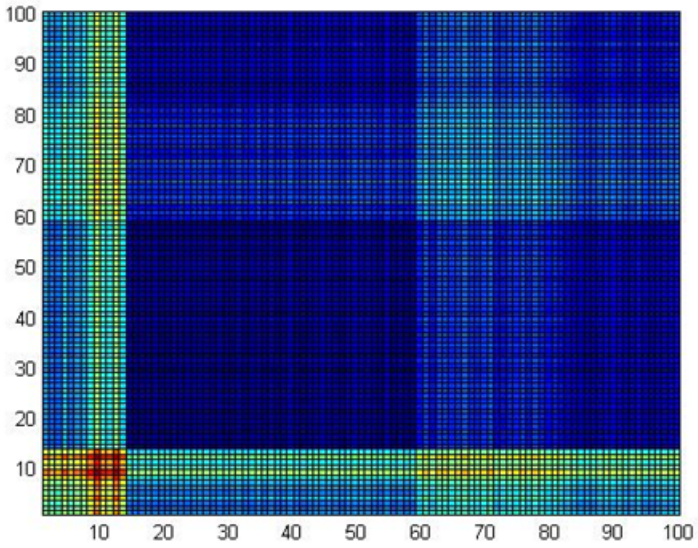


Figura 29: Matriz de Afinidade para agrupamentos formados com valor de k = 3 com algoritmo Kernel K-Médias na base de dados TV News Commercial Detection.

A matriz para o algoritmo Kernel K-Medias também mostra uma grande similaridade entre os grupos obtidos, apesar de ser um grau menor do que verificado para o algoritmo K-Médias. Isso se dá pelos valores de silhouette encontrados, e disposição dos dados na base. Comparando com a matriz verificada anteriormente temos bons resultados, com pequenas quantidades de similaridade e uma matriz diagonal que é predominantemente coesa.

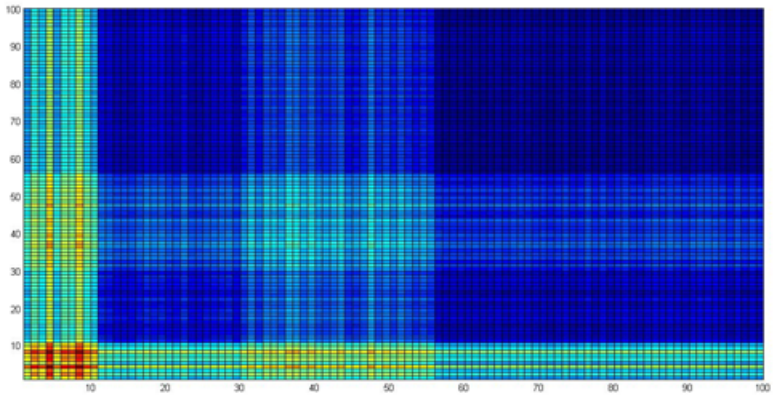


Figura 30: Matriz de Afinidade para agrupamentos formados com valor de k = 3 com algoritmo K-Medóides na base de dados TV News Commercial Detection.

O algoritmo K-Medóides tem uma matriz que tem valores de coesão mais claros, denotando uma similaridade não tão grande. Isso é resultado das médias silhouette que o algoritmo apresentou. Apesar disto neste agrupamento em particular temos valores ainda aceitáveis que mostram um agrupamento razoavelmente compacto dentre as opções calculadas nos testes para o algoritmo.

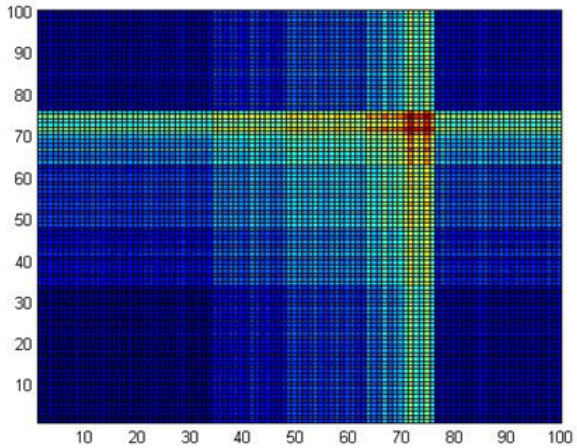


Figura 31: Matriz de Afinidade para agrupamentos formados com valor de k = 8 com algoritmo Fuzzy C-Médias na base de dados TV News Commercial Detection.

A matriz de afinidade gerada para o algoritmo Fuzzy C-Médias está com grupos bastante coesos, sendo apenas uma exceção. Este valor de agrupamentos obteve uma ótima média de silhouette, sendo próxima do maior valor encontrado. Este algoritmo se mostra um bom exemplo onde a base ficou bem espaçada e dividida em um número maior de agrupamentos como é visto na matriz da figura 25.

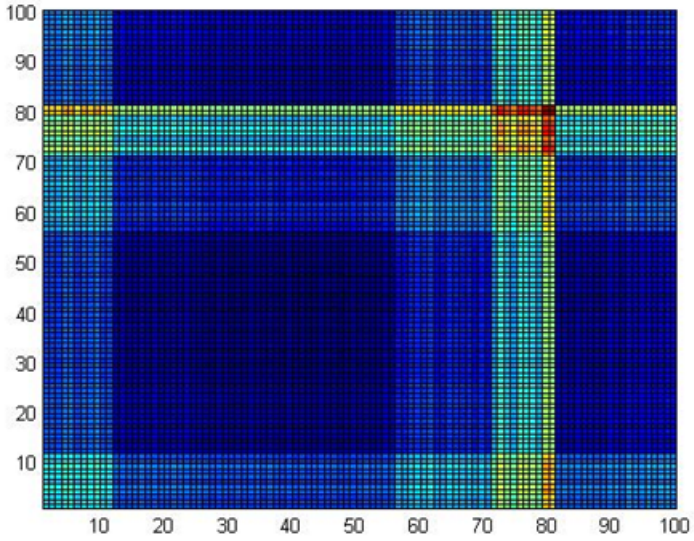


Figura 32: Matriz de Afinidade para agrupamentos formados com valor de k = 7 com algoritmo Spectral Clustering na base de dados TV News Commercial Detection.

A matriz de afinidade para o algoritmo *Spectral Clustering*, assim como a matriz de afinidade para os agrupamentos da Figura 21, mostra um alto grau de coesão. Com valores mostrados, a maioria dos grupos se mostra de boa qualidade, com alguma exceção. A matriz diagonal tem uma cor predominante, e isso mostra a compactividade e demonstração da qualidade dos grupos gerados pelo algoritmo.Mais uma vez com um número maior de grupos, a base ficou visivelmente bem dividida.

Para comparação de desempenho de algoritmos em relação ao tempo de execução, ou seja, o tempo em que obtemos uma resposta acerca do melhor valor de k, assim tendo o número de agrupamentos que melhor se encaixa ao padrão dos dados, a Figura 27 nos dá os melhores tempos de cada algoritmo.

Tabela 6: Valores de tempo de execução para melhores agrupamentos na base TV News

Algoritmo	**Tempo de Execução**
K-Médias	12,4 horas
K-Medóides	0,30 horas
Kernel K-Médias	0,50 horas
Fuzzy C-Médias	12,3 horas
Spectral Clustering	2,12 horas

Vimos que o K-Medóides obteve o resultado no menor tempo, sendo acompanhado de perto pelo algoritmo Kernel K-Médias. Spectral Clustering obteve um valor mediano e o restante dos algoritmos continuou com tempos bastante altos. Assim como aconteceu para a base de dados HHAR, os algoritmos Fuzzy C-Médias e K-Médias tiverem tempos grandes, porém para os melhores agrupamentos a média de silhouette foi boa. K-Medóides ainda continuou sendo o algoritmo mais rápido tanto para a base HHAR, quanto para a base TV News. É notado também uma evolução do algoritmo *Spectral Clustering*, que obteve uma diminuição de tempo de execução de cerca de três horas. O algoritmo Kernel K-Médias também obteve diminuição do tempo, fazendo com que as respostas fossem dadas cerca de duas horas mais rápido.

6.3 Testes com Base de Dados Youtube Multiview Video Games

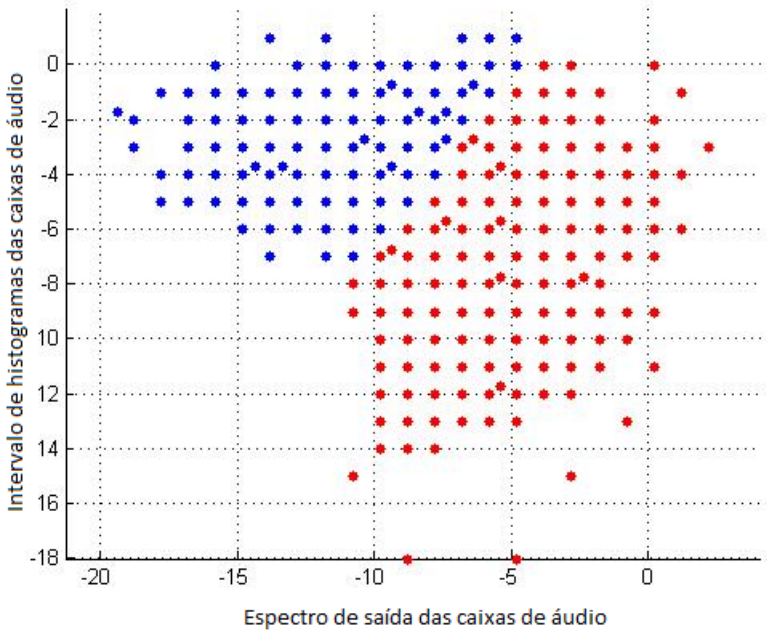
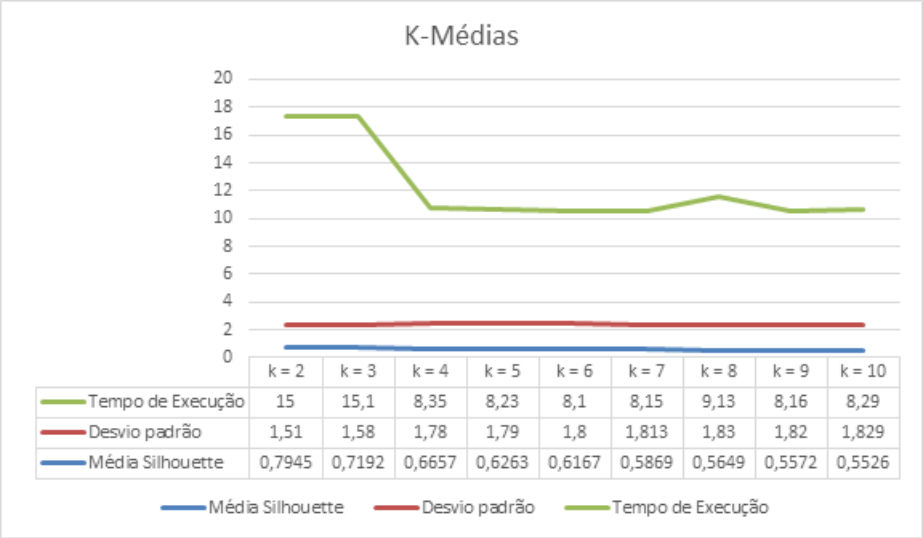


Figura 33: Representação gráfica dos agrupamentos formados pelo algoritmo K-Médias na base de dados YMVG.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	15	15,1	8,35	8,23	8,1	8,15	9,13	8,16	8,29
Desvio padrão	1,51	1,58	1,78	1,79	1,8	1,813	1,83	1,82	1,829
Média Silhouette	0,7945	0,7192	0,6657	0,6263	0,6167	0,5869	0,5649	0,5572	0,5526

Figura 34: Resultados obtidos para algoritmo K-Médias para base de dados YMVG.

Na base de dados YMVG repetimos todos os testes com os mesmos algoritmos. Nesta base o número de atributos é maior, porém as instâncias são diminuídas ainda mais. Apesar disso o algoritmo K-Médias apresentou o mesmo comportamento de pouca variação na média silhouette, dando uma qualidade homogênea independentemente do valor de agrupamentos. Sendo o melhor valor com um número de dois grupos. O desvio padrão também segue o mesmo comportamento, assim como na base de dados anterior. O que mais chamou a atenção foi que para agrupamentos pequenos, dois e três respectivamente, tivemos um tempo de execução superior ao visto nos testes anteriores, porém ao aumentar o valor de k, o tempo foi abaixando com valores consideráveis, dando respostas em até cinco horas e meia mais rápidas.

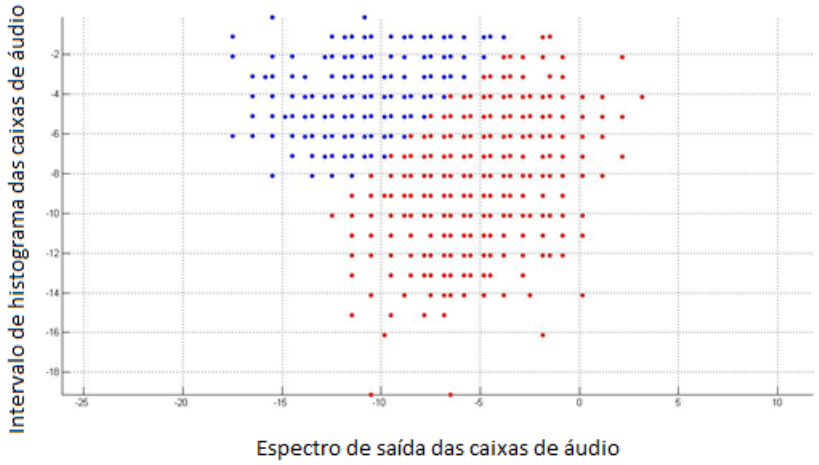
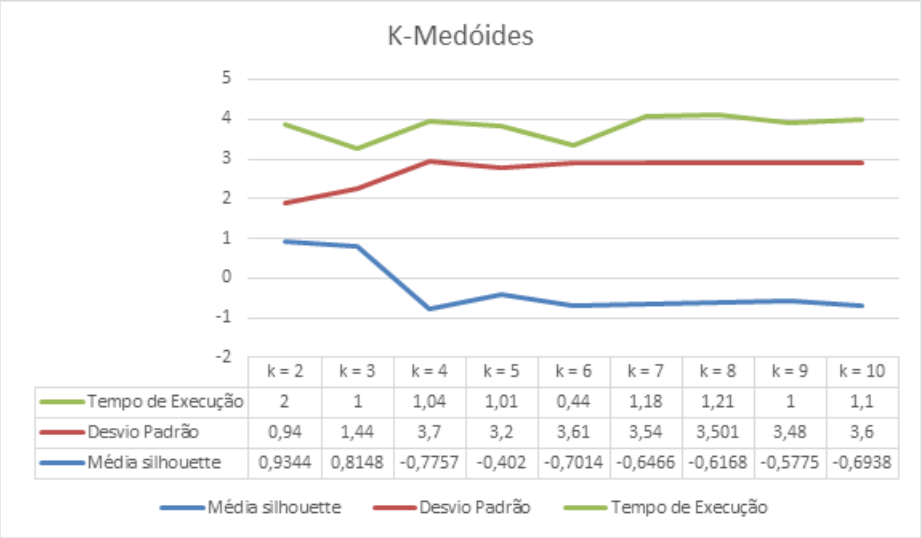


Figura 35: Representação gráfica dos agrupamentos formados pelo algoritmo K-Medóides na base de dados YMVG.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	2	1	1,04	1,01	0,44	1,18	1,21	1	1,1
Desvio Padrão	0,94	1,44	3,7	3,2	3,61	3,54	3,501	3,48	3,6
Média silhouette	0,9344	0,8148	-0,7757	-0,402	-0,7014	-0,6466	-0,6168	-0,5775	-0,6938

Figura 36: Resultados obtidos para algoritmo K-Medóides para base de dados YMVG.

O algoritmo k-Medóides demonstrou grande capacidade nos testes com a base de dados YMVG. Obtendo o maior valor de média silhouette dentre todos os testes feitos. Chegando próximo do valor de similaridade total dentre os grupos criados, para o melhor caso. Porém, a medida que se aumentava o número de k, o algoritmo perdia qualidade. Tanto é, que a partir de um número de agrupamento tivemos os piores valores registrados nos testes feitos para a similaridade dos grupos. Para esse algoritmo os dados se apresentavam essencialmente em um padrão de dois a três agrupamentos somente. Qualquer valor diferente destes se mostrava altamente dissemelhante. Também foi notado um aumento do tempo de execução do algoritmo que nas bases anteriores se mostrou o mais eficaz em função do tempo da resposta obtida. Curiosamente a perda de desempenho da média silhouette aconteceu a partir do mesmo instante onde o algoritmo K-Médias registrou melhora no tempo de execução.

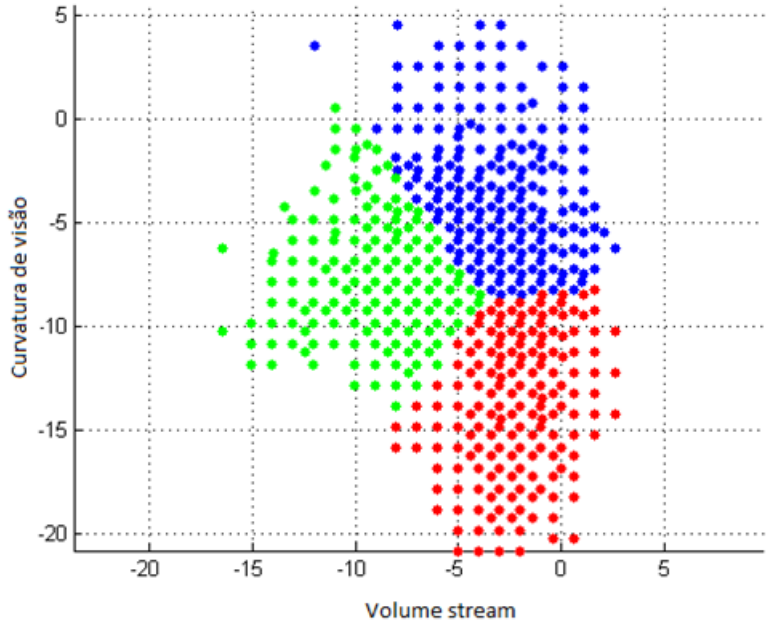
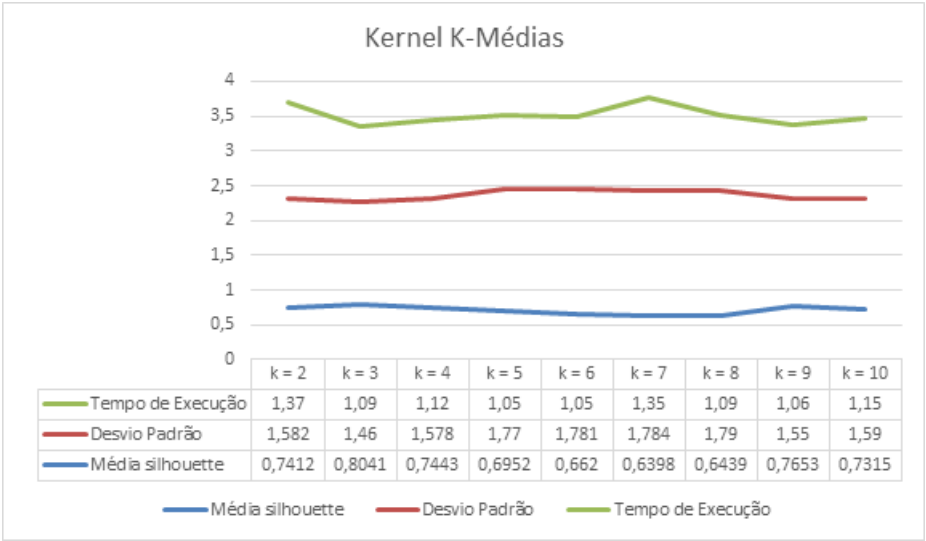


Figura 37: Representação gráfica dos agrupamentos formados pelo algoritmo Kernel K-Médias na base de dados YMVG.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	1,37	1,09	1,12	1,05	1,05	1,35	1,09	1,06	1,15
Desvio Padrão	1,582	1,46	1,578	1,77	1,781	1,784	1,79	1,55	1,59
Média silhouette	0,7412	0,8041	0,7443	0,6952	0,662	0,6398	0,6439	0,7653	0,7315

Figura 38: Resultados obtidos para algoritmo Kernel K-Médias para base de dados YMVG.

O algoritmo Kernel k-Médias se diferenciou dos algoritmos anteriores para a base de dados YMVG, pelo fato de ter achado um número diferente de agrupamentos ótimos. Para este algoritmo o melhor valor de k é igual a três. Os outros algoritmos tinham melhores resultados para um valor de k igual a dois. Este algoritmo se mostrou bastante eficaz com valores sempre acima de 0,6 para todos os valores de k testados. A média silhouette se mostrou alta e com isso a qualidade dos agrupamentos era verificada. O desvio padrão também se mostrou pouco variável, tendo diferenças sutis para testes diferentes. O tempo

de execução também foi aceitável, se distorcendo em apenas dois casos onde demorou mais de uma hora e meia para gerar a saída desejada.

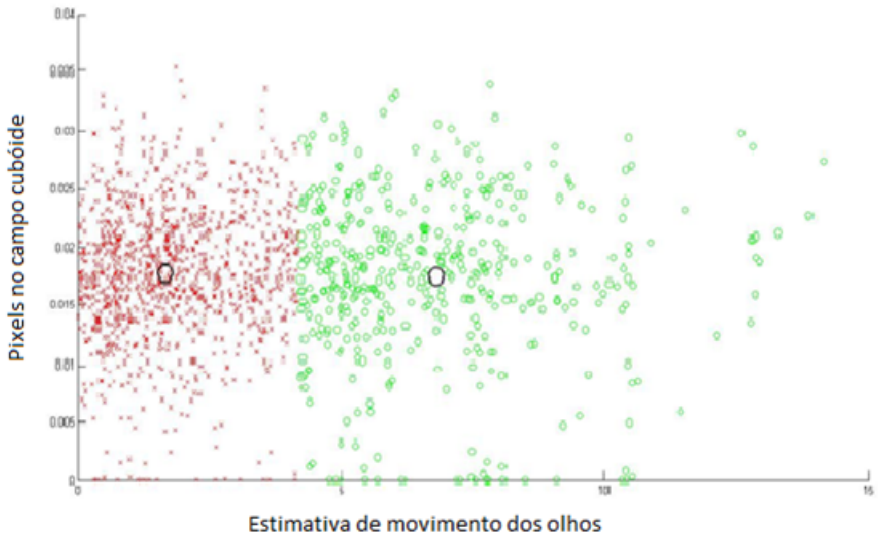
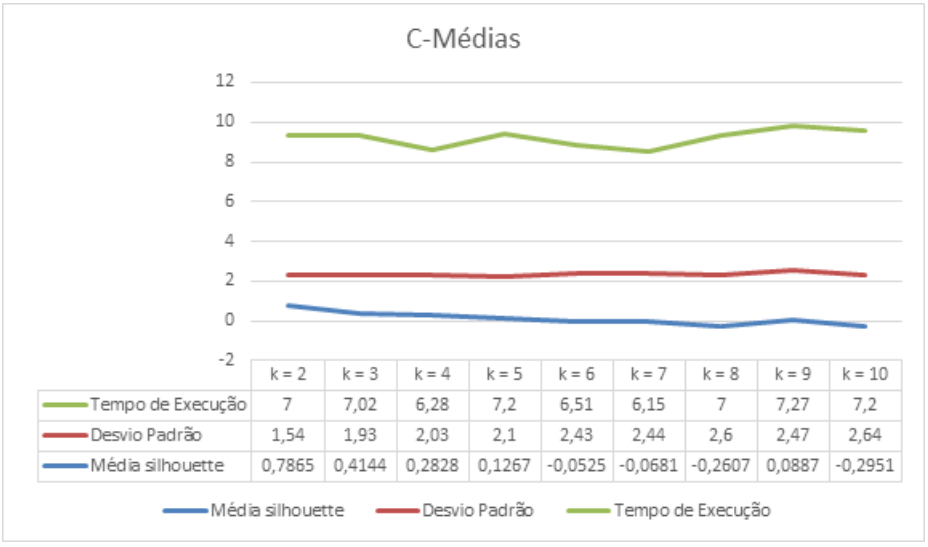


Figura 39: Representação gráfica dos agrupamentos formados pelo algoritmo Fuzzy C-Médias na base de dados YMVG.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	7	7,02	6,28	7,2	6,51	6,15	7	7,27	7,2
Desvio Padrão	1,54	1,93	2,03	2,1	2,43	2,44	2,6	2,47	2,64
Média silhouette	0,7865	0,4144	0,2828	0,1267	-0,0525	-0,0681	-0,2607	0,0887	-0,2951

Figura 40: Resultados obtidos para algoritmo Fuzzy C-Médias para base de dados YMVG.

O algoritmo Fuzzy C-Médias obteve valores que alternaram de muito bons para ruins. No melhor caso, para um valor de k igual a dois, a média foi muito boa, com alto grau de coesão. Porém em alguns casos, o algoritmo se mostrou ineficiente, criando agrupamentos com valores silhouette negativos, mostrando haver similaridade no mesmo. Os valores de desvio padrão se mostraram bons no melhor caso, porém foram crescendo à medida que se aumentava a quantidade de agrupamentos. O ponto positivo para a base de dados YMVG foi o tempo de execução do algoritmo. Se mostrou o mais eficiente nesse quesito dentre todos os testes feitos para o C-Médias. E se mostrou pouco variável para

os valores diferentes de k. Levando a uma generalização do tempo independente desta variável.

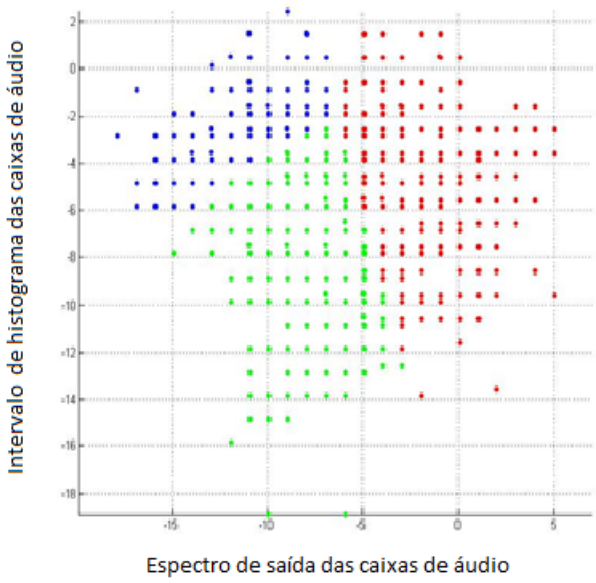
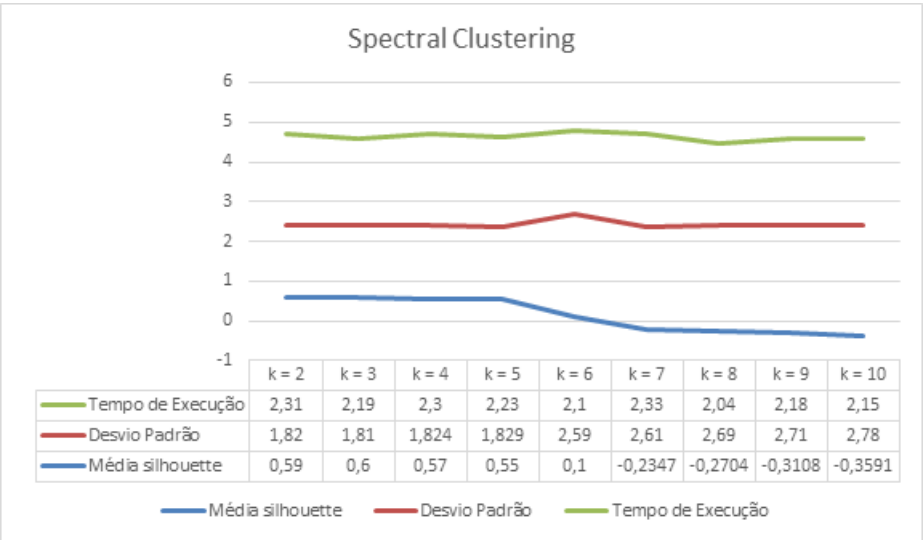


Figura 41: Representação gráfica dos agrupamentos formados com algoritmo Spectral Clustering na base de dados YMVG.



	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8	k = 9	k = 10
Tempo de Execução	2,31	2,19	2,3	2,23	2,1	2,33	2,04	2,18	2,15
Desvio Padrão	1,82	1,81	1,824	1,829	2,59	2,61	2,69	2,71	2,78
Média silhouette	0,59	0,6	0,57	0,55	0,1	-0,2347	-0,2704	-0,3108	-0,3591

Figura 42: Resultados obtidos para algoritmo Spectral Clustering para base de dados YMVG.

O algoritmo *Spectral Clustering* apresentou um mau desempenho para a base de dados YMVG. Com valores baixos para os agrupamentos, e a medida em que era aumentado este valor, a média silhouette abaixava para valores que mostravam dissimilaridade nos grupos criados. Assim como a média silhouette, o desvio padrão aumentava seu valor à medida que o número de grupos crescia. Mostrando perda de desempenho e distanciamento dos dados dentro do agrupamento. O tempo de execução se mostrou estável para todos

os valores de k, não variando conforme os outros valores. Em média o algoritmo nos apresentava as respostas em duas horas e poucos minutos de diferença de um caso a outro.

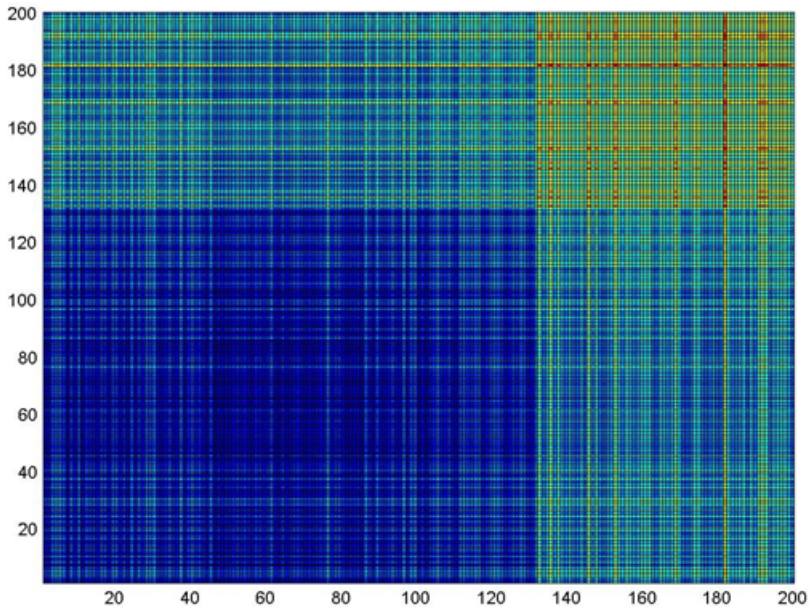


Figura 43: Matriz de Afinidade para agrupamentos formados com valor de k = 2 com algoritmo K-Médias na base de dados YMVG.

A matriz de afinidade obtida ao utilizar os agrupamentos gerados com o algoritmo K-Médias tem uma boa similaridade e coesão. Isso é mostrado devido a cor, quase homogênea da matriz. Esta é altamente compacta, com valores que se assemelham bem em seus grupos, e nos grupos adjacentes.

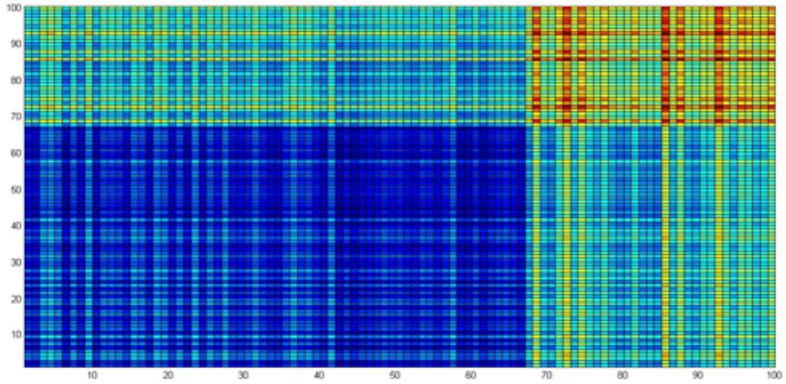


Figura 44: Matriz de Afinidade para agrupamentos formados com valor de k = 2 com algoritmo Kernel k-Médias na base de dados YMVG.

A matriz de afinidade para os agrupamentos gerados com o algoritmo Kernel K-Médias se mostra menos coesa do que a anterior, para o algoritmo K-Médias. Isto é

devido a variação entre os grupos, e os valores menores de média silhouette. No caso onde temos dois agrupamentos vemos uma diagonal, menos coesa mais ainda assim com um valor aceitável e que é considerado bom para as dimensões da base de dados.

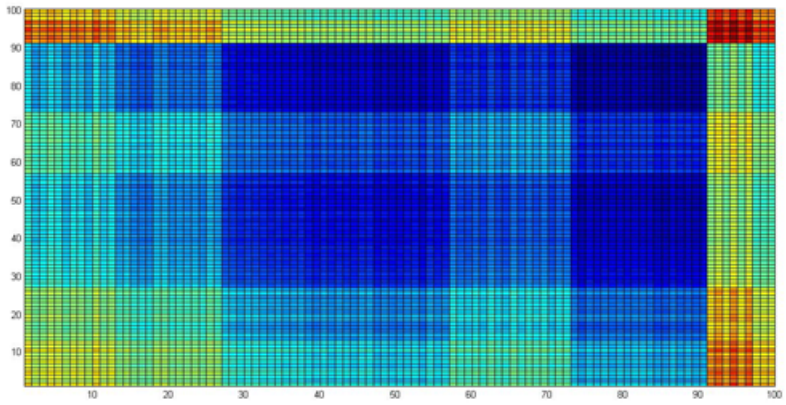


Figura 45: Matriz de Afinidade para agrupamentos formados com valor de k = 10 com algoritmo K-Medóides na base de dados YMVG.

O algoritmo K-Medóides se mostrou altamente variável nos testes à medida que aumentávamos o número de agrupamentos a ser criado. Com isso a matriz de afinidade apresenta grande grau de dissimilaridades. Mostrando uma pequena coesão e pouca compatibilidade entre os dados de um mesmo grupo. Isso é verificado pela predominância de cores fortes na matriz principal. E pela matriz em si se mostrar bastante dividida nas demais partes.

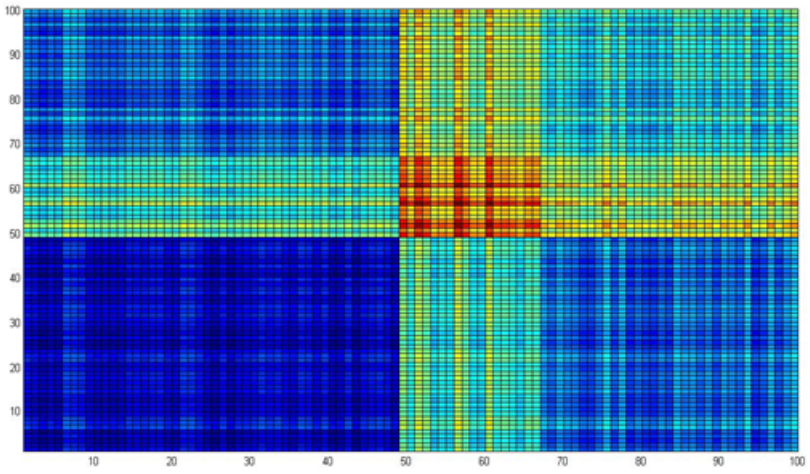


Figura 46: Matriz de Afinidade para agrupamentos formados com valor de k = 3 com algoritmo Fuzzy C-Médias na base de dados YMVG.

No algoritmo Fuzzy C-Médias temos uma matriz de afinidade onde predomina-se os grupos coesos, onde os valores silhouette tendem a ser maiores. A divisão mostra que

a maioria dos valores dissimilares estão nas diagonais adjacentes, e dentro da diagonal principal temos uma área menos coesa, mas que não muda a qualidade dos grupos, como pode ser visto na tabela 19, que contém as médias de silhouette para o algoritmo.

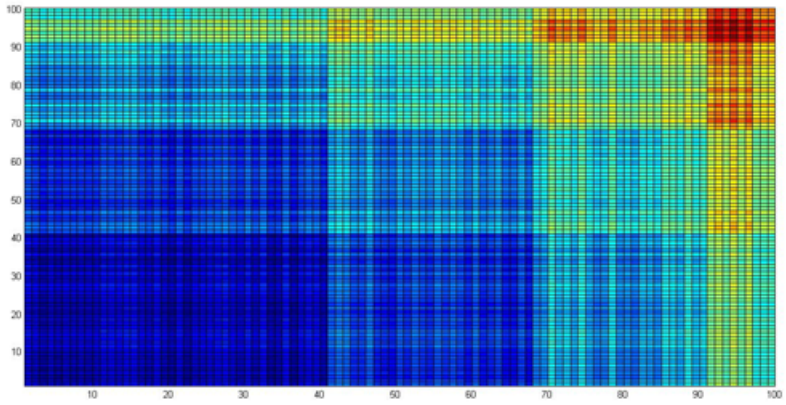


Figura 47: Matriz de Afinidade para agrupamentos formados com valor de k = 7 com algoritmo Spectral Clustering na base de dados YMVG.

A matriz de afinidade para o algoritmo *Spectral Clustering* tem uma coesão grande para os valores menores e à medida que é aumentado os valores a dissimilaridade se mostra presente. Com uma parte com cores fortes, mostra uma pequena coesão, mesmo sendo uma pequena área mas que interfere na parte principal da diagonal. E como foi notado pelos valores da Figura 41 o algoritmo não obteve valores muito promissores para a base de dados YMVG.

Tabela 7: Valores de tempo de execução para melhores agrupamentos na base YMVG.

Algoritmo	**Tempo de Execução**
K-Médias	15 horas
K-Medóides	2 horas
Kernel K-Médias	1,09 horas
Fuzzy C-Médias	7 horas
Spectral Clustering	2,19 horas

A Tabela 7 apresenta o tempo de execução para cada algoritmo usado nos testes, na descoberta do número ótimo de agrupamentos. Mesmo alguns algoritmos achando valores diferentes para k, é possível fazer essa comparação devido à pouca diferença entre esses valores. Os algoritmos K-Médias, K-Medoides e Kernel K-Médias apresentaram aumento desse tempo, em relação as outras bases de dados. Sendo que, o algoritmo K-Medóides foi o que apresentou o maior aumento desse tempo, e ao mesmo tempo que conseguiu obter a melhor média para o valor silhouette, também teve as piores médias para agrupamentos diferentes do melhor caso. O algoritmo Fuzzy C-Médias foi aquele que teve a maior melhora nesse quesito, com ganho de cinco horas em geral para gerar a resposta. O desempenho

também foi maior na base de dados YMVG. Spectral Clustering também apresentou melhora de desempenho no quesito tempo de resposta, mas não apresentando variação na qualidade dos agrupamentos.

Conclusão e Trabalhos Futuros

Baseando-se nas mesmas métricas de avaliação temos que todos os algoritmos conseguiram desempenhar o papel de encontrar padrões para destacar os agrupamentos, sendo que alguns o realizaram com três grupos enquanto outros consideraram apenas dois. Outros valores de agrupamentos esporadicamente apresentavam um valor elevado, mas nenhum padrão pode ser tirado desses valores.

Dentre todos os utilizados aquele que apresentou o pior desempenho foi o Fuzzy C-Médias, tendo o pior valor de validação, neste estudo, utilizando o método silhouette. Também tendo um valor médio de desvio padrão maior, apesar de que, este número pode ser compreendido devido a uma quantidade gigantesca de dados, que variam entre números pequenos, negativos e positivos, e valores mais expressivos.

O algoritmo C-Médias junto com o K-Médias foram os que apresentaram o maior tempo de execução nas três bases de dados, chegando a tempos de 11 horas a mais do que os demais algoritmos em alguns casos.

O algoritmo K-Medóides foi o que apresentou os menores tempos de execução, superando os demais algoritmos em cerca de 3 a 40 por cento, dependendo da quantidade de grupos, variação do algoritmo e base de dados. Em alguns casos o algoritmo Kernel K-Médias também apresentou valores menores que os demais, mas no geral o desempenho ainda foi satisfatório.

K-Medóides foi também aquele que mais teve perda de desempenho à medida que aumentávamos o número de agrupamentos, isso mostra como para este algoritmo os padrões dos dados não suportavam números maiores de grupos, além dos necessários. Para a base de dados YMVG o número de agrupamentos superior a quatro mostrava grande grau de dissimilaridade e pequena coesão. Obtendo os piores valores dentre todos os testes entre todas as bases. Curiosamente nesse mesmo caso, tivemos o maior valor de média de silhouette.

Em média o algoritmo *Spectral Clustering* apresentou bons valores, tendo um tempo computacional razoável, baseando-se nos demais e com valores maiores para a validação. Isso demonstrou a coesão e similaridade com maior grau para este algoritmo. O desvio padrão também foi um ponto positivo para este método. Esse desempenho superior é visto com maior facilidade na base de dados TV News, onde para um tempo computacional, não tão grande a média de silhouette dos agrupamentos foram bem satisfatórios.

O objetivo de analisar o comportamento e desempenho destes algoritmos nas grandes bases de dados foi satisfeito, pois foi comprovado por meio de experimentos,

seus pontos positivos e negativos, e por meio dos valores mostrados temos uma métrica consistente capaz de nos dizer qual algoritmo trabalha melhor para as bases de dados trabalhadas.

E como era esperado o desempenho destes algoritmos cai à medida que a base de dados se torna maior e mais complexa. Com a adição de mais dimensões tempos um custo maior, que demanda mais tempo e isso reflete na qualidade dos grupos e do cálculo da distância entre os dados.

É deixado como um fator a ser analisado futuramente, como o desempenho é mais afetado pela adição de atributos na base de dados do que pela adição de instancias. Como foi mostrado nos resultados obtidos o uso de dados com mais atributos e menos dados causa uma variação maior na obtenção do melhor valor de agrupamentos. Ao se achar este valor, levando mais tempo de execução, os valores tendem a ser mais elevados, porém os valores fora deste padrão "ótimo" tendem a ser menores do que aqueles usados em bases maiores porém com menos atributos.

O que pode ser deixado como contribuição é a necessidade de melhoria para estes métodos, ou mesmo o uso de diferentes técnicas para melhorar o tempo de execução gasto para bases de dados ainda maiores que possam precisar ser agrupadas. Também encontrar o equilíbrio entre o desempenho, o custo, e a qualidade dos grupos para que os padrões corretos possam ser encontrados e utilizados posteriormente para realizar algum tipo de classificação ou mineração dos dados.

Como trabalho futuro deve ser dado mais atenção aos valores muito discrepantes. Utilizar mais métricas de avaliação de distâncias para verificar a magnitude dos valores dos atributos. O nível de correlação dos dados também deve ser verificado, e sugere-se uma avaliação inicial abrangente para que as amostras sejam identificadas e tome-se medidas de condições de falhas. Em (JANGARELLI et al., 2015) sugere-se reduzir o número de dimensões, e até mesmo reduzir a sensibilidade de valores iniciais altamente correlativos.

Referências

ARORA, S.; CHANA, I. A survey of clustering techniques for big data analysis. In: IEEE. *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-*. [S.l.], 2014. p. 59–65. Citado 3 vezes nas páginas 32, 34 e 43.

CALON, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature genetics*, Nature Publishing Group, v. 47, n. 4, p. 320–329, 2015. Citado na página 33.

CHARRAD, M. *STDHA*. 2016. Disponível em: <http://www.sthda.com/english/wiki/clustering-validation-statistics-4-vital-things-everyone-should-know-unsupervised-machine-learning#silhouette-analysis>. Acesso em: 31 apr 2016. Citado na página 41.

COSTA, J. A. F.; NETTO, M. L. D. A. Estimating the number of clusters in multivariate data by self-organizing maps. *International Journal of Neural Systems*, World Scientific, v. 9, n. 03, p. 195–202, 1999. Citado na página 31.

DEMCHENKO, Y.; LAAT, C. D.; MEMBREY, P. Defining architecture components of the big data ecosystem. In: IEEE. *Collaboration Technologies and Systems (CTS), 2014 International Conference on*. [S.l.], 2014. p. 104–112. Citado 3 vezes nas páginas 24, 25 e 43.

DHILLON, I. S.; GUAN, Y.; KULIS, B. Kernel k-means: spectral clustering and normalized cuts. In: ACM. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2004. p. 551–556. Citado 2 vezes nas páginas 34 e 35.

DUARTE, F. J. F. *Optimização da Combinação de Agrupamentos baseado na Acumulação de Provas pesadas por Índices de Validação e com uso de Amostragem*. Tese (Doutorado) — Universidade de Trás-os-Montes e Alto Douro, 2008. Citado 8 vezes nas páginas 19, 29, 35, 36, 38, 40, 43 e 44.

FRANÇA, A. S. d. Otimização do Processo de Aprendizagem da Estrutura Gráfica de Redes Bayesianas em BigData. *UFPA-Universidade do Guamá - Belém/Pará*, n. 20/02, 2014. Disponível em: <http://repositorio.ufpa.br/jspui/handle/2011/5608>. Citado na página 43.

HAMAD, D.; BIELA, P. Introduction to spectral clustering. In: *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*. [S.l.: s.n.], 2008. Citado 2 vezes nas páginas 29 e 37.

JAIN, A. K. Clustering Big Data. *Department of Computer Science Michigan State University*, p. 1–44, 2012. Disponível em: <http://www.cse.nd.edu/Fu{_}Prize{_}Seminars/jain/slid>. Citado 3 vezes nas páginas 19, 29 e 31.

JANGARELLI, M. et al. Análise de agrupamento de diferentes densidades de marcadores no mapeamento genético por varredura genômica. *Ceres*, v. 57, n. 6, 2015. Citado na página 80.

LANGONE, R. et al. Kernel spectral clustering and applications. In: *Unsupervised Learning Algorithms*. [S.l.]: Springer, 2016. p. 135–161. Citado na página 29.

LELE, C.; MOUTARI, S. Computational methods for study of foldness of h-ideals in bci-algebras. *Soft computing*, Springer, v. 12, n. 4, p. 403–407, 2008. Citado na página 31.

LOPES, P. de A. Agrupamento de dados semi-supervisionado no contexto de aprendizado de máquina. 2009. Citado 5 vezes nas páginas 30, 31, 34, 37 e 40.

MADANI, O.; GEORG, M.; ROSS, D. A. On using nearly-independent feature families for high precision and confidence. *Machine learning*, Springer Science+ Business Media, Van Godewijckstraat 30 Dordrecht 3311 GX Netherlands, v. 92, n. 2-3, p. 457–477, 2013. Citado na página 28.

MAXWELL, I. E. *Managing sustainable innovation: The driver for global growth*. [S.l.]: Springer Science & Business Media, 2009. Citado 4 vezes nas páginas 30, 33, 34 e 36.

NG, A. Y. et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, MIT; 1998, v. 2, p. 849–856, 2002. Citado na página 34.

QUEIROZ, F. A. de A. Um estudo sobre métodos de kernel para classificação e agrupamento de dados. UFMG, 2009. Citado na página 33.

RAMOS, A. Estudos de técnicas de transferência de aprendizado para mineração de dados. 2014. Citado 2 vezes nas páginas 11 e 44.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987. Citado 2 vezes nas páginas 24 e 41.

SHINNOU, H.; SASAKI, M. Spectral clustering for a large data set by reducing the similarity matrix size. In: *LREC*. [S.l.: s.n.], 2008. Citado 2 vezes nas páginas 37 e 38.

STISEN, A. et al. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In: ACM. *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. [S.l.], 2015. p. 127–140. Citado 2 vezes nas páginas 25 e 26.

SWAN, M. Philosophy of big data: Expanding the human-data relation with big data science services. In: IEEE. *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*. [S.l.], 2015. p. 468–477. Citado na página 25.

VALENTE, D. X. R. Uma avaliação da utilização de matrizes de afinidades na validação de agrupamentos de dados. 2013. Citado 3 vezes nas páginas 11, 42 e 44.

VYAS, A. et al. Commercial block detection in broadcast news videos. In: ACM. *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. [S.l.], 2014. p. 63. Citado na página 27.

WU, X. et al. Data mining with big data. *IEEE transactions on knowledge and data engineering*, IEEE, v. 26, n. 1, p. 97–107, 2014. Citado 2 vezes nas páginas 19 e 24.

ZHANG, R.; RUDNICKY, A. I. A large scale clustering scheme for kernel k-means. In: IEEE. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. [S.l.], 2002. v. 4, p. 289–292. Citado na página 34.