Clustering Techniques for Profile Construction Using 5G Access Network Data

Yves Dantas Neves[1], **Lázaro Raimundo de Oliveira**[1], **João Carlos Xavier Júnior**[1], **Anne M.P. Canuto**[2]

[1]Instituto Metrópole Digital - Universidade Federal do Rio Grande do Norte Natal, RN - Brasil

[2]Departamento de Informática e Matemática Aplicada - Universidade Federal do Rio Grande do Norte, Natal, RN - Brasil

yvesneves@gmail.com, lazaro.oliveira.813@ufrn.edu.br,

jcxavier@imd.ufrn, anne.canuto@ufrn.br

Abstract. *The Internet and the development of Information and Communication Technologies have increased the volume and diversity of data sources, opening new opportunities in sectors for the application of Machine Learning and Big Data technologies. In this perspective, the Mobile Network Access infrastructure has been generating extensive amount of data. The main aim of this work is to apply clustering algorithms in order to identify profiles from data generated by 5G access network indicators regarding traffic, volume and channel quality. From the clustering partitions, we use the data profiles as dataset for a classification tool aiming to support fault identification, performance management and operational efficiency of access networks.*

Resumo. *A Internet e o desenvolvimento das Tecnologias da Informação e Comunicação expandiram o volume e a diversidade das fontes de dados, abrindo assim novas oportunidades nos setores industriais e acadêmicos à aplicações de tecnologias relacionadas ao Aprendizado de Máquina (AM) e Big Data. Nessa perspectiva, a infraestrutura de Acesso das Redes Móveis vem gerando uma grande quantidade de dados. O presente trabalho tem como objetivo principal a aplicação de algoritmos de agrupamento para a criação de perfis a partir de dados relacionados à indicadores de redes de acesso 5G referentes a tráfego, volume e qualidade de canal. A partir das partições geradas na etapa de agrupamento, espera-se que os perfis encontrados possam servir de base para uma ferramenta de suporte a identificação de falhas, gestão de desempenho e eficiência operacional das redes de acesso.*

1. Introdução

Segundo [Reed 2012], uma habilidade cognitiva básica do ser humano é agrupar informações semelhantes para identificar padrões e classificar objetos e eventos. Assim, o reconhecimento de padrões surge como um estágio inicial do processamento de informações, relacionado à memorização e aprendizagem.

O reconhecimento de padrões, aliado ao uso de técnicas de Aprendizado de Máquina (AM), tem ganhado destaque no setor de Telecomunicações. A expectativa é

que os algoritmos de AM desempenhem um papel crucial na operação eficiente das redes 5G, auxiliando na gestão do grande volume de dados gerado e fornecendo suporte à tomada de decisões [Chiu et al. 2017].

A aplicação de técnicas não supervisionadas de AM em dados de desempenho permite identificar padrões de tráfego em células, ajudando a definir parâmetros de configuração e identificar células com comportamentos semelhantes. Isso busca melhorar a eficácia na resolução de problemas na rede [Raivio et al. 2003]. Técnicas de agrupamento podem ser usadas para encontrar células com desempenho similar em termos de qualidade de sinal, taxas de queda de chamada e volume de tráfego [Wang and Ferrús 2021].

Por último, além de auxiliar na detecção de problemas como erros de configuração e falhas nos equipamentos de rede, a análise de agrupamento baseada em perfis de desempenho pode facilitar a seleção de conjuntos de dados (datasets) para treinar algoritmos supervisionados voltados à predição de tráfego e planejamento de recursos [Zhu and Sun 2020]. Assim, este artigo propõe o uso de técnicas de agrupamento para construir perfis de desempenho a partir de indicadores como taxa de transferência, volume e qualidade de canal em redes 5G. Os conjuntos de dados serão escolhidos com base em métricas da literatura para garantir sua adequação a problemas de classificação de tráfego e desempenho.

O restante deste artigo está dividido em 6 seções. A Seção 2 descreve alguns conceitos teóricos importantes, já na seção 3 é mostrado alguns trabalhos relacionados ao tema. Enquanto a seção 4 apresenta a metodologia utilizada no desenvolvimento deste trabalho, ao passo que uma análise dos resultados é mostrada na Seção 5 e, finalmente, a Seção 6 apresenta as considerações finais a respeito deste trabalho.

2. Referencial Teórico

2.1. Redes Móveis

Um sistema de redes móveis (ou rede celular) é composto por uma cadeia de transmissão, na qual a conexão final é realizada sem a utilização de cabeamento, dispersa em células fornecidas por estações bases. Uma célula (elemento fundamental na composição de uma rede móvel) consiste em uma área geográfica que define uma zona de cobertura do serviço fornecido por uma estação base (também conhecida como torre de celular) a partir de transceptores que transmitem e recebem sinais de rádio em frequências licenciadas para os telefones celulares dos usuários [Angadi et al. 2019].

As redes utilizam indicadores de desempenho ou KPI's (do inglês *Key Performance Indicators*) como forma de medir e avaliar a qualidade dos sistemas e a experiência dos usuários. A obtenção dos indicadores de desempenho é realizada a partir da coleta de medições do tráfego das células em operação e a avaliação estatística das medições coletadas são utilizadas para a tomada de decisão sobre iniciativas que otimizem a utilização dos recursos de rádio disponíveis. O tipo de KPI pode indicar onde a rede precisa ser ajustada [Angadi et al. 2019].

2.2. Aprendizado de Máquina

Há uma grande variedade de técnicas de Aprendizado de Máquina encontradas na Literatura. Em geral, essas técnicas são classificadas em três tipos: (i) Aprendizado não

Supervisionado que é utilizado em tarefas de descrição onde as informações relevantes são identificadas em dados sem a presença de um elemento externo para guiar o aprendizado; (ii) Aprendizado Supervisionado que é utilizado em tarefas de previsão onde o objetivo principal é a definição de um modelo a partir dos dados de treinamento de forma que possa ser utilizado na previsão de um rótulo ou valor que caracterize um novo exemplo com base nos valores de seus atributos de entrada e de saída utilizando técnicas de classificação ou regressão; (iii) Aprendizado Semissupervisionado que é utilizado em tarefas de previsão e que consiste em uma intersecção das abordagens supervisionadas e não supervisionadas onde os rótulos estão presentes em apenas uma parte pequena dos dados de treinamento [Faceli et al. 2021].

Entre as técnicas de Aprendizado não supervisando de Máquina, as técnicas de Agrupamento ou *Clustering* buscam agrupar pontos de dados (instâncias) não rotulados em um certo número de grupos com o intuito de explorar estruturas implícitas e fornecer informações úteis para análises mais avançadas. Visando analisar uma boa diversidade de técnicas, para este trabalho, escolheu-se as seguintes: (i) k-Means [Everitt et al. 2009] - baseada em distância; (ii) DBSCAN [Aggarwal and Reddy 2018] - baseada em densidade; (iii) Expectation Maximization (EM) [Aggarwal and Reddy 2018] - baseada em distribuição. O algoritmo EM utiliza o GMM (*Gaussian mixture model*) como modelo de distribuição normal.

No contexto do Aprendizado Supervisionado, se destacam as técnicas de Classificação e Regressão. Porém, nesse estudo não foram utilizadas técnicas de regressão. Dessa forma, dentre as técnicas de classificação utilizadas sobre as partições criadas na fase de agrupamento estão as seguintes: k-NN (k-Nearest Neighbor), MLP (MultiLayer Parceptron) e AD (Árvore de Decisão) [Faceli et al. 2021].

2.3. Critérios de Validação

Os Critérios de Validação consistem nas abordagens utilizadas para validação de um agrupamento. Essas abordagens utilizadas para a comparação entre diferentes técnicas ou para determinação de um valor mais apropriado em algum parâmetro referente ao algoritmo aplicado. Dentre os vários índices utilizados em critérios de validação, os mais comuns são o Índice Dunn [Dunn 2008], o Davies-Bouldin [Davies and Bouldin 1979] e o Silhouette [Rousseeuw 1987].

3. Trabalhos Relacionados

3.1. Utilização de Técnicas de AM sobre dados de Redes Móveis

Baseado nos trabalhos encontrados, de forma geral, é possível constatar que diversos autores têm focado suas pesquisas no uso de técnicas de AM no intuito de resolver alguns problemas, tais como: predição de requerimentos inerentes aos sistemas de redes móveis (por exemplo, latência excessivamente agressivas, regimes de conectividade massiva com baixo consumo de energia e baixo esforço computacional) [Srinivasan et al. 2023], [Chen et al. 2020]; e planejamento de infraestrutura de redes (por exemplo, posicionamento das estações base e suas interconexões regionais) [Guo and Zhang 2022].

Além desses trabalhos, outros focaram especificamente na análise de problemas de desempenho e investigação de falhas [Raivio et al. 2003], [Santos et al. 2019]. O propósito de ambos era analisar perfis de desempenho e configurações aplicadas em redes

móveis através da aplicação de técnicas de agrupamento. Em um outro trabalho, os autores buscavam a detecção de anomalias nas redes também através do uso de *Clustering* [Liu et al. 2019].

Por último, vários trabalhos investigaram padrões de tráfego em estações bases [Xu et al. 2017], [Margaris et al. 2022], [Jun et al. 2013], [Parwez et al. 2017], [Hashmi et al. 2017] e [Rekkas et al. 2021]. As pesquisas concentraram na aplicação de algoritmos de agrupamento para identificação de padrões e otimização de indicadores relacionados à intensidade do sinal de referência recebido, relação sinal-ruído e a taxa de transferência. Todos esses trabalhos buscavam padrões para identificação de melhorias na alocação de recursos e aplicação de soluções para mitigação de falhas.

3.2. Discussão

No que se refere aos trabalhos relatados, percebe-se que na grande maioria deles as aplicações estavam relacionadas à otimização de capacidade da rede e a detecção de anomalias no tráfego. Além desses dois temas, a segurança e privacidade em redes móveis também foi tópico de interesse, principalmente em [Guo and Zhang 2022] e [Srinivasan et al. 2023]. Por outro lado, outros trabalhos focaram na aplicação de técnicas de agrupamento para encontrar perfis relacionados ao desempenho em redes de acesso (por exemplo, indicadores de desempenho). O que diferenciou tais trabalhos foi a escolha dos algoritmos de agrupamento.

Finalmente, diferentemente dos trabalhos que utilizam técnicas de Aprendizado de Máquina (supervisionado e não supervisionado), este trabalho propõe a utilização de várias técnicas de agrupamento sobre dados oriundos de redes móveis, tais como: retenção, disponibilidade, volume, taxa de transferência e interferência. Dessa forma, através da análise de agrupamento, será possível gerar e selecionar as melhores partições, para a partir dessas (transformação em datasets), fazer predições que permitirão a tomada de decisão relacionada aos indicadores de desempenho. Por fim, todos os datasets gerados na etapa de agrupamento serão avaliados na ótica da predição, mais especificamente classificação, através da utilização de três técnicas de classificação.

4. Metodologia Experimental

A presente seção descreve a metodologia experimental realizada no intuito de avaliar o desempenho dos algoritmos de agrupamento (DBSCAN, GMM e K-Means) na busca por padrões (perfis) em células com desempenho semelhantes. Como pode ser visualizado na Figura 1, os dados são coletados, depois os atributos são selecionados, e na sequência os algoritmos de agrupamento são utilizados sobre os dados pré-processados. Como resultado, temos as partições que são selecionadas a partir dos critérios de validação, para finalmente termos os datasets que serão utilizados na etapa de classificação (aplicação das técnicas KNN, MLP e AD).

4.1. Coleta de dados

Depois de coletados, os dados foram fatiados em três porções: (i) uma porção reduzida com 4.144 amostras, (ii) uma porção intermediária com 8.288 amostras, e (iii) o conjunto de dados completo com 12.432 amostras. É importante salientar que os dados são coletados em intervalos de 15 minutos a partir de 296 células previamente selecionadas

Figura 1. Fluxograma Metodológico

durante um período de 42 semanas. Dessa forma, foram coletados 12.432 registros de desempenho das células.

4.2. Atributos Selecionados

Para o experimento, foram selecionados dados referentes a qualidade de canal, taxa de transferência e volume de tráfego das células. No que se refere aos indicadores selecionados, foram considerados um atributo de acessibilidade relacionado a disponibilidade (razão de disponibilidade temporal das células), um índice de utilização relacionado ao volume do tráfego dos usuários e um índice de integridade referente a taxa de transmissão o qual considera a razão entre o volume de dados transmitidos e o tempo acumulado pelas representações dos bits nos sinais físicos (realizados pelo esquema de modulação do 5G).

Além dos já citados, foram também considerados indicadores relacionados à interferência de *uplink*, volume de *uplink* e volume de *downlink*. Por último, também foram considerados os atributos de acessibilidade e retenção, e de forma sumarizada, a Tabela 1 lista todos os atributos selecionados para o procedimento experimental, além de suas respectivas unidades de medida. Esses indicadores sugerem melhorias referentes a cobertura, capacidade e refinamento de configurações de mobilidade.

A seleção dos atributos também permite identificar grupos de células em estado de mal funcionamento, conhecidas como Sleepy Cells, onde a falha não aciona alarmes e a comunicação na área fica inacessível. Essas falhas podem durar dias antes de serem detectadas [Manzanilla-Salazar et al. 2020]. Para lidar com a variabilidade e as diferentes unidades de medida, os dados foram normalizados antes da aplicação dos algoritmos.

Variável	Medida
Retenção	%
Disponibilidade	%
Volume Uplink	MB
Volume Downlink	MB
Taxa de Transferência	Mbps
Interferência PUCCH	dBm
Interferência PUSCH	dBm

Tabela 1. Indicadores de Desempenho de Rede de Acesso 5G: Retenção (expressa em percentual); Disponibilidade (expressa em percentual); Volume UpLink e DownLink (expressa em Megabytes); Taxa de TRansferência (expressa em Mega bits por segundo); Interferência PUCCH e PUSCCH (expressa em Decibéis miliwatts - dBm).

4.3. Aplicação de Técnicas de Aprendizado de Máquina

Dentro do escopo de análise de dados de redes de acesso e considerando as técnicas disponíveis, a aplicação de algoritmos de agrupamento tem sido utilizada como um método de análise simples, eficaz e que tem atraído a atenção de muitos pesquisadores [Wang and Ferrús 2021]. Portanto, os algoritmos escolhidos para serem utilizados sobre os três datasets foram os seguintes: (i) DBSCAN (baseado em densidade), (ii) Gaussian Mixture Models (baseado em distribuição) e (iii) o K-Means (baseada em distância).

No que se refere a utilização do algoritmo DBSCAN, os dois principais parâmetros (MinPts [2-52] e Eps [0.1-0.9]) foram ajustados através do método *Grid Search*. Dessa forma, o referido algoritmo foi executado n vezes (sendo n o tamanho do espaço de busca) sobre cada um dos três datasets, e por último, cada partição gerada foi avaliada pelo índice de validação Dunn. Ao final, baseado nos valores apresentados pelo índice Dunn, escolheu-se os valores para o DBSCAN, sendo MinPts = 9 e Eps = 0,8.

Por outro lado, como os algoritmos EM (GMM) e k-Means necessitam do valor de k a priori, aplicou-se o método do Cotovelo (*Elbow Method*) para identificar os valores de k mais apropriados segundo o índice de validação. Inicialmente, para essa finalidade, os valores de k variaram entre 2-20.

Na sequência, como fruto da etapa de agrupamento, avaliou-se todas as partições criadas pelos três algoritmos sobre a ótica dos índices de validação DB e Silhouete. Baseado nos dois índices, selecionou-se as partições com melhores valores de índice e com número de grupos semelhantes, para que, dessa forma, sejam realizados experimentos com os algoritmos de classificação (AD, k-NN e MLP).

Os três algoritmos de classificação tiveram seus principais parâmetros ajustados de acordo com o valor padrão de cada um. Cada configuração (algoritmo + parâmetros) foi treinada e testada sobre a ótica da validação cruzada (k-fold cv) com o valor de k igual a 10. Utilizou-se a métrica da Acurácia para avaliar o desempenho dos três modelos.

Os resultados dos algoritmos foram avaliados pelo teste de Friedman, amplamente utilizado em aprendizado de máquina para comparar o desempenho de múltiplos algoritmos em diferentes bases de dados. Se o teste indicar uma diferença estatística significativa, um teste post-hoc pode ser aplicado para identificar quais algoritmos se destacam estatisticamente [Demšar 2006].

5. Resultados Experimentais

É importante ressaltar que os três algoritmos de agrupamento (k-Means, DBSCAN e EM-GMM) foram executados sobre três conjuntos de dados diferentes, intitulados de porção reduzida, porção intermediária e porção completa. O primeiro compreende 14 semanas de coleta de dados com 4.144 instâncias, o segundo compreende 28 semanas com 8.288 instâncias, e o terceiro que compreende 42 semanas com 12.432 instâncias.

5.1. Experimentos com o Agoritmo DBSCAN

A execução do DBSCAN para os três conjuntos de dados produziu partições com características semelhantes. Para cada um dos conjuntos de dados, foi detectado um grupo de instâncias ruidosas contendo atributos com valores extremamente discrepantes das médias

Conjuntos	Índice Davies-Bouldin	Índice Silhouette	No de Grupos
Reduzido	2,030766384	0,243894844	5
Intermediário	**1,771322227**	0,523671886	5
Completo	1,782742443	**0,563059406**	4

Tabela 2. Índices de Validação para os Experimentos com DBSCAN

dos atributos pertencentes aos conjuntos de dados originais. Respectivamente, as porções reduzida, intermediária e completa apresentaram 193, 275 e 328 amostras anômalas.

No que se refere aos Índices de Validação e aos grupos gerados (partições), como pode ser visualizado na Tabela 2, é possível notar que os melhores valores para ambos os índices ocorreram quando o algoritmo foi executado sobre o conjunto Completo (por exemplo, 1,7827 e 0,5630). Além disso, o algoritmo gerou apenas quatro grupos, diferindo dos demais conjuntos onde foram gerados cinco grupos.

Por último, se faz necessário enfatizar que para todos os conjuntos, sempre houve um grupo com instâncias ruidosas, algo comum quando se utiliza esse tipo de algoritmo de agrupamento. Dessa forma, os grupos a serem considerados são: reduzido (4), Intermediário (4) e Completo (3).

Usando DBSCAN foram identificadas as prováveis associações aos clusters: grupo (1) degradação de acessibilidade; grupo (2) degradação de disponibilidade e baixo volume de dados; grupo (3) degradação de disponibilidade e elevada interferência no uplink; grupo (4) degradação na taxa de transferência e elevada interferência no uplink e; (5) grupo de uso médio.

5.2. Experimentos com o Algoritmo EM

O Algoritmo *Expectation Maximization* (usando um GMM) foi executado sobre os três conjuntos de dados, variando os valores de k entre 2 e 20. Dessa forma, avaliou-se os melhores valores para o índice Dunn. Para os três conjuntos, o melhor valor foi verificado para $k = 6$.

A Tabela 3 apresenta os valores encontrados para os índices de validação Davies-Bouldin e Silhouette. Note que os valores do índice Silhouette foram baixos para todos os conjuntos, se comparados com os valores do experimento anterior. Além disso, ambos os índices foram unânimes em apontar o conjunto reduzido como o melhor entre os três. Finalmente, é importante ressaltar que os conjuntos Intermediário e Completo apresentaram muita heterogeneidade, ou seja, baixo valor (+/- 2%) para o índice Silhouette.

Usando EM foram identificadas as prováveis associações aos clusters: grupo (1) alto volume de tráfego; grupo (2) degradação da taxa de transferência e volume de dados abaixo da média; grupo (3) alta interferência no uplink; grupo (4) degradação nos indicadores de acessibilidade e retenção e; grupo (5) taxa de transferência acima da média e volume acima da média.

5.3. Experimentos com o Algoritmo K-Means

O Algoritmo k-Means foi executado sobre os três conjuntos de dados, variando os valores de k entre 2 e 20. Dessa forma, avaliou-se os melhores valores para o índice Dunn. Para os três conjuntos, o melhor valor foi verificado para $k = 5$.

Conjuntos	Índice Davies-Bouldin	Índice Silhouette
Reduzido	1,94051421	**0,243894844**
Intermediário	**2,52696936**	0,01310830
Completo	2,27164967	0,01736042

Tabela 3. Índices de Validação para os Experimentos com EM com k = 6

Conjuntos	Índice Davies-Bouldin	Índice Silhouette
Reduzido	0,70851055	0,38681724
Intermediário	**0,68796621**	**0,38881748**
Completo	1,06709714	0,24824980

Tabela 4. Índices de Validação para os Experimentos com k-Means com k = 5

No que se refere aos valores encontrados para os índices de validação Davies-Bouldin e Silhouette, apresentados na Tabela 4, pode-se afirmar que os valores do índice DB foram os melhores para todos os três conjuntos se comparados com os experimentos anteriores. Por outro lado, com relação ao índice Silhouette, pode-se afirmar que para os conjuntos Reduzido e Intermediário tal índice obteve cerca de 38% de homogeneidade, perdendo apenas para os valores encontrados para o DBSCAN que chegaram a 52% e 56% para os conjuntos Intermediário e Completo, respectivamente.

Usando K-Means foram identificadas as prováveis associações aos clusters: grupo (1) baixo volume de tráfego; grupo (2) degradação no indicador de disponibilidade; grupo (3) interferência de uplink acima da média; grupo (4) degradação nos indicadores de acessibilidade e retenção e; grupo (5) taxa de transferência e volume acima da média.

5.4. Escolha das Melhores Partições

A escolha das melhores partições foi baseada em fatores relacionados aos bons resultados do índice de validação de Silhouette que representa homogeneidade (ver Tabela 5). Além disso, levou-se em consideração as partições com o mesmo número de grupos, sendo cinco grupos nesses caso. Por último, considerou-se as técnicas de agrupamento, ou seja, DBSCAN e k-Means.

No que se refere aos índices de validação, os valores encontrados para os conjuntos Reduzido e Intermediário representaram os melhores resultados para todos os índices calculados, como pode ser visto na Tabela 5.

Técnica	Conjunto	Índice Silhouette
DBSCAN	Reduzido	0,24389484
DBSCAN	Intermediário	**0,52367188**
k-Means	Reduzido	0,38681724
k-Means	Intermediário	**0,38881748**

Tabela 5. Melhores partições geradas pelo DBSCAN e K-means com 5 grupos.

5.5. Experimentos com Técnicas de Classificação

Depois de escolhidas as melhores partições geradas pelos algoritmos de Agrupamento, se fez necessário conduzir uma análise empírica no que se refere a acurácia dos algoritmos de classificação, sendo eles: Árvore de Decisão (AD), k-NN (k Nearest Neighbors) e MultiLayer Perceptron (MLP). Nesse caso, uma alta acurácia pode significar que os grupos gerados pelos algoritmos de agrupamento realmente representam perfis bem robustos e fácies de serem classificados.

Como apresentado na seção de metodologia experimental, para a realização dos experimentos de classificação e verificação de acurácia, foi utilizado o método k-Fold Cross Validation, sendo *k* igual a 10 para que fosse realizada uma validação cruzada nos resultados. Na Tabela 6 é possível visualizar os valores da acurácia para os experimentos realizados em cada um dos algoritmos.

Conjuntos de Dados	Treinamento/Teste	k-NN	AD	MLP
Metodologia		Acc	Acc	Acc
DBSCAN - Reduzido	10-fold CV	0,984	0,965	0,978
DBSCAN - Intermediário	10-fold CV	0,991	0,978	0,985
k-Means - Reduzido	10-fold CV	0,976	0,919	0,993
k-Means - Intermediário - Reduzido	10-fold CV	0,980	0,929	0,996
	Média ⇒	0,983	0,948	0,988
	Desv. Pad. ⇒	0,006	0,028	0,008

Tabela 6. Acurácia das Classificações Baseadas nas Melhores Rotulações.

Como pode ser visualizado na Tabela 6, os valores de Acurácia obtidos por cada um dos algoritmos de classificação mostram que a classificação das bases (partições) originárias da etapa de agrupamento conseguiu identificar de forma fácil os perfis de indicadores presentes nos dados. Isso pode ser comprovado através da acurácia média e do desvio padrão de cada algoritmo de classificação (duas últimas linhas da Tabela 6).

Analisando individualmente os algoritmos de classificação, é possível notar que o algoritmo MLP obteve a maior acurácia média (Acc_media = 0,988), ficando o k-NN com a segunda melhor (Acc_media = 0,983) e a AD em terceiro (Acc_media = 0,948). Também é preciso ressaltar que o desvio padrão das acurácias foi muito pequeno, sendo 0,006 para o k-NN, 0,028 para a AD e 0,008 para o MLP. Dessa forma, como já esperado, o teste estatístico de Friedman não apontou diferença estatística entre as acurácias (p-value = 0,41686).

A respeito das quatro partições (DBSCAN-Reduzido, DBSCAN-Intermediário, k-Means-Reduzido e k-Means-Intermediário), é possível ver que a acurácia média de cada base, em função do desempenho dos três classificadores, variou entre 94% e 98,5%. Além disso, as duas primeiras obtiveram 97,5% e 98,5% de acurácia média, respectivamente. Contudo, não houve diferença estatística entre as acurácias (p-value = 0,47529).

Apenas como forma de verificação e comparação do desempenho dos algoritmos preditivos utilizando as partições com os piores índices de validação, a Tabela 7 apresenta a acurácia dos algoritmos executados sobre todas as partições produzidas pelo algoritmo EM. Como é possível notar, a acurácia média diminuiu consideravelmente. Porém, de

Conjuntos de Dados	Treinamento/Teste	k-NN	AD	MLP
Metodologia		Acc	Acc	Acc
EM - Reduzido	10-fold CV	0,755	0,835	0,924
EM - Intermediário	10-fold CV	0,696	0,796	0,918
EM - Completo	10-fold CV	0,678	0,778	0,917
	Média ⇒	0,710	0,803	0,920
	Desv. Pad. ⇒	0,040	0,029	0,004

Tabela 7. Acurácia das Classificações Baseadas nas Piores Rotulações.

forma parecida com o primeiro experimento, o algoritmo MLP obteve a maior média (0,920). O algoritmo AD obteve a segunda melhor média (0,803), ficando o k-NN em último (0,710).

6. Considerações Finais

Este trabalho apresentou uma abordagem baseada em algoritmos de agrupamento (DBSCAN, EM e k-Means) capaz de identificar e extrair padrões (indicadores de desempenho) em dados oriundos de redes móveis 5G. A partir dela, os padrões podem ser analisados, facilitando a correção dos indicadores de desempenho.

No que se refere aos grupos gerados pelos algoritmos de agrupamento, foi possível notar que o DBSCAN identificou instâncias com ruido onde as células apresentaram desempenho muito divergentes como, por exemplo, pico abrupto de volume (o que pode ser resultado de uma eventual mudança no comportamento dos usuários) e índice de acessibilidade próximo a zero (o que pode caracterizar uma Sleepy Cell). Por outro lado, os algoritmos EM e k-Means foram capazes identificar um grupo de células com indicadores de acessibilidade e retenção zerados, demonstrando sua efetividade na detecção de possíveis Sleepy Cells.

Também é importante destacar que o DBSCAN foi capaz de rotular um grupo de instâncias com degradação no indicador de acessibilidade. A formação de grupos com características homogêneas - que objetiva a utilização das análises de agrupamento - é interessante pois segmenta conjuntos de células em que planos de melhoria podem ser aplicados de forma unificada. Além disso, ambos EM e k-Means foram capazes de identificar grupos de células com perfis de desempenho acima da média. A identificação de células com bom desempenho é útil para a realização de análises mais avançadas referentes aos atributos e a configuração de parâmetros aplicados às células, para que, dessa forma, sejam utilizadas como base para aplicação de melhorias em outras células com desempenho degradado.

Outro fator importante identificado nas partições geradas está relacionado aos perfis associados ao volume de tráfego acima da média e ao volume abaixo da média. A identificação desse tipo de perfil é importante pois o acompanhamento dos indicadores de utilização (como o de volume) e sua associação com os indicadores de acessibilidade, retenção e taxa de transferência são importantes para que haja um planejamento mais efetivo da distribuição de recursos e uma previsão nos investimentos necessários para expensão do atendimento aos usuários.

Por fim, no que se refere aos algoritmos de classificação, a avaliação dos resul-

tados de acurácia demonstra que os dados rotulados facilitaram a tarefa preditiva. Se considerados os valores resultantes de acurácia, todos os experimentos com as melhores partições apresentaram valores superiores a 90% (acurácia média). Dessa forma, os experimentos com os modelos preditivos apresentaram mais uma comprovação de que os perfis encontrados nas partições geradas foram robustos e bem definidos. Consequentemente, também, demonstrando viabilidade do uso de técnicas de aprendizado não supervisionado para classificação do tráfego das redes e posterior uso como variável resposta em técnicas supervisionadas, possibilitando redução de custos em implementações reais.

Como trabalho futuro, pretende-se expandir a nossa abordagem como forma de analisar especificamente as anomalias presentes nos dados, favorecendo a construção de um sistema automático para detecção de anomalias em tráfego de redes móveis 5G.

Referências

Aggarwal, C. and Reddy, C. (2018). *Data Clustering: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press.

Angadi, A. V., Yashaswini, S. D., Balaji, K., and Padma, U. (2019). Rf planning and optimization practices applied to improve the kpi's of 4g lte network. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 848–851.

Chen, Y., Liu, W., Niu, Z., Feng, Z., Hu, Q., and Jiang, T. (2020). Pervasive intelligent endogenous 6g wireless systems: Prospects, theories and key technologies. *Digital Communications and Networks*, 6(3):312–320.

Chiu, P., Reunanen, J., Luostari, R., and Holma, H. (2017). Big data analytics for 4.9g and 5g mobile network optimization. In *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pages 1–4.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

Dunn, J. (2008). Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4:95–104.

Everitt, B. S., Landau, S., and Leese, M. (2009). *Cluster Analysis*. Wiley Publishing, 4th edition.

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. d., and Carvalho, A. C. P. d. L. F. d. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

Guo, R. and Zhang, J. (2022). Research on 5g communication station location planning and regional clustering based on k-medoids and dbscan algorithm. In *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, pages 1097–1101.

Hashmi, U. S., Darbandi, A., and Imran, A. (2017). Enabling proactive self-healing by data mining network failure logs. In *2017 International Conference on Computing, Networking and Communications (ICNC)*, pages 511–517.

Jun, L., Tingting, L., Gang, C., Hua, Y., and Zhenming, L. (2013). Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis. *China Communications*, 10(12):25–36.

Liu, X., Chuai, G., Gao, W., Zhang, K., and Chen, X. (2019). Kqis-driven qoe anomaly detection and root cause analysis in cellular networks. In *2019 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6.

Manzanilla-Salazar, O. G., Malandra, F., Mellah, H., Wetté, C., and Sansò, B. (2020). A machine learning framework for sleeping cell detection in a smart-city iot telecommunications infrastructure. *IEEE Access*, 8:61213–61225.

Margaris, A., Filippas, I., and Tsagkaris, K. (2022). Hybrid network–spatial clustering for optimizing 5g mobile networks. *Applied Sciences*, 12(3).

Parwez, M. S., Rawat, D. B., and Garuba, M. (2017). Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network. *IEEE Transactions on Industrial Informatics*, 13(4):2058–2065.

Raivio, K., Simula, O., Laiho, J., and Lehtimaki, P. (2003). Analysis of mobile radio access network using the self-organizing map. In *IFIP/IEEE Eighth International Symposium on Integrated Network Management, 2003.*, pages 439–451.

Reed, S. (2012). *Cognition: Theories and Applications*. Cengage Learning.

Rekkas, V. P., Sotiroudis, S., Sarigiannidis, P., Karagiannidis, G. K., and Goudos, S. K. (2021). Unsupervised machine learning in 6g networks -state-of-the-art and future trends. In *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–4.

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

Santos, R., Sousa, M., Vieira, P., Queluz, M. P., and Rodrigues, A. (2019). An unsupervised learning approach for performance and configuration optimization of 4g networks. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6.

Srinivasan, M., Skaperas, S., Mitev, M., Herfeh, M. S., Shehzad, M. K., Sehier, P., and Chorti, A. (2023). Smart channel state information pre-processing for authentication and symmetric key distillation. *IEEE Transactions on Machine Learning in Communications and Networking*, 1:328–345.

Wang, S. and Ferrús, R. (2021). Extracting cell patterns from high-dimensional radio network performance datasets using self-organizing maps and k-means clustering. *IEEE Access*, 9:42045–42058.

Xu, F., Li, Y., Wang, H., Zhang, P., and Jin, D. (2017). Understanding mobile traffic patterns of large scale cellular towers in urban environment. *IEEE/ACM Transactions on Networking*, 25(2):1147–1161.

Zhu, Q. and Sun, L. (2020). Big data driven anomaly detection for cellular networks. *IEEE Access*, 8:31398–31408.