Análise de Incidentes de Data Center através da Aplicação de Técnica de Mineração de Dados

Data Center Incident Analysis through the Application of Data Mining Technique

Samuel Luna Martins[1]
orcid.org/0000-0002-1464-3782

Carmelo Bastos Filho[1]
orcid.org/0000-0002-0924-5341

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.

E-mail: samuellunamartins@gmail.com; carmelo.filho@upe.br

DOI: 10.25286/repa.v7i2.2221

Como citar este artigo pela NBR 6023/2018: Samuel Luna Martins Carmelo Bastos Filho. Análise de Incidentes de Data Center através da Aplicação de Técnica de Mineração de Dados. Revista de Engenharia e Pesquisa Aplicada, v.7, n. 2, p. 86-95, 2022.

RESUMO

Esse trabalho tem como objetivo aplicar a técnica de agrupamento K-Modes a fim de auxiliar na identificação das causas-raízes dos problemas de disponibilidade e desempenho de serviços e sistemas hospedados em servidores e máquinas virtuais de um Data Center de uma organização. Os dados foram extraídos a partir da ferramenta de monitoramento chamada Zabbix relativos aos últimos 3 meses de incidentes. Foi realizado um procedimento de pré-processamento dos dados, extraindo os atributos mais relevantes, posteriormente foi aplicada a técnica chamada K-Modes juntamente com o valor de K mais adequado encontrado a partir do método Elbow. Após análise de dados, foi possível extrair regras de correlação e criar um plano estratégico a fim de mitigar a quantidade de incidentes recorrentes.

PALAVRAS-CHAVE: Incidentes; Mineração de Dados; Network Operation Center; Data Center; Zabbix.

ABSTACT

This work aims to apply the K-Modes grouping technique in order to help identify the root causes of problems related to the availability and performance of services and systems hosted on servers and virtual machines in an organization's Data Center. Data were extracted from the monitoring tool called Zabbix regarding the last 3 months of incidents. A data pre-processing procedure was performed, extracting the most relevant attributes, then the technique called K-Modes was applied along with the most suitable K value found from the Elbow method. After analyzing the data, it was possible to extract correlation rules and create a strategic plan in order to mitigate the amount of recurring incidents.

KEY-WORDS: Incident; Data Mining; Network Operation Center; Data Center; Zabbix

1 INTRODUÇÃO

Em ambientes de Data Centers complexos que possuem diversos ativos de TI, uma ferramenta de monitoramento de incidentes pode em algumas situações registrar milhares de registros em um curto espaço de tempo. É comum se verificar que alguns eventos são recorrentes e que, na verdade, podem pertencer a um problema maior.

Em um cenário em que o número de especialistas de TI em uma organização é limitado e a quantidade de eventos de incidentes no Data Center supera a capacidade de atuação em tempo hábil, uma pergunta que emerge é: como identificar quais são os incidentes ou tipos de incidentes mais relevantes para colocarmos como meta de resolução para o próximo mês baseado nos meses anteriores? Focar apenas nos incidentes de severidade alta, que normalmente causam problemas de indisponibilidade, nem sempre é o adequado para se achar a causa-raiz de problemas.

Algumas das ferramentas de monitoramento consolidadas e presentes nas organizações possuem diversos recursos, inclusive de predição. Porém, a triagem de incidentes é algo ainda desafiador e que não é suportado nativamente por várias dessas ferramentas.

Desta forma, faz-se necessário o uso de um método inteligente capaz de analisar milhares de registros de logs de monitoramento a fim de apontar quais são os incidentes ou tipos de incidentes que mais comprometem o bom funcionamento do NOC de uma organização. A partir desses agrupamentos de incidentes mais relevantes, realizar uma análise mais profunda para encontrar uma correlação entre eles a fim de achar uma ou mais causas-raiz para os problemas.

Ao lançar olhos sobre a área de mineração de dados em ciências da computação, percebe-se que a aplicação de uma técnica de agrupamento de dados exportados a partir de uma ferramenta de monitoramento e com o devido pré-processamento poderá revelar quais grupos de incidentes são os mais relevantes. Essa técnica poderá auxiliar também na correlação dos eventos a fim de achar as causas-raiz dos problemas a fim de minimizar os incidentes mais recorrentes e que mais impactam na qualidade de monitoramento de Data Centers.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentadas as principais tecnologias e abordagens para realização deste trabalho.

2.1 NETWORK OPERATION CENTER

Network Operation Center (NOC) é um local em que se monitora, supervisiona, delega e soluciona incidentes de Data Center dos mais variados tipos e severidades [1]. O NOC é o ponto focal para solução de problemas de infraestrutura de TI. Ambientes de Data Centers possuem muitos ativos de TI como servidores, máquinas virtuais e dispositivos de rede que podem ocorrer falhas e gerar inúmeros registros de incidentes.

2.2 FERRAMENTA DE MONITORAMENTO DE DATA CENTER

Sistemas de monitoramento de Data Center são ferramentas auxiliares que monitoram os ativos de TI e geram alertas para a equipe de monitoramento quando algo não vai bem na infraestrutura. Existem diversas soluções, sejam gratuitas e de código-fonte aberto e outras proprietárias que podem possuir de forma intrínseca mecanismos inteligentes de auxílio na solução de problemas. Zabbix é um sistema de monitoramento de código aberto para Data Centers que fornece métricas de monitoramento baseados em modelos [2]. Desde o lançamento em 1998, vem sendo adotada mundialmente com mais de 300 mil instalações e pelos órgãos públicos do Brasil como é o caso do SERPRO [3].

2.3 TÉCNICA DE MINERAÇÃO DE DADOS K-MODES

O algoritmo K-Modes é utilizado para agrupar dados categóricos. A estratégia baseia-se na definição de grupos com base no número de categorias correspondentes entre os pontos de dados. Diferentemente do algoritmo K-Means que agrupa dados numéricos com base na distância euclidiana [4].

O K-Modes trabalha com dados estruturados. Um Sistema de Informação Categórica também conhecido como uma Tabela Categórica armazena em cada linha em uma tupla que representam fatos sobre um objeto [5] [6]. Uma informação

DOI: 10.25286/repa.v7i2.2221

categórica pode ser descrita como $Inf = (U, A, V, f)$, onde [7]:

1. U é o conjunto não vazio de objetos, que pode ser chamado de universo.
2. A é o conjunto não vazio de atributos.
3. V é a união do domínio de atributos, por exemplo: $V = \cup_{a \in A} V_a$ onde V_a é o valor de domínio de atributo a e é finito e não ordenado.
4. $f: UxA \rightarrow V$ é a função informação que para cada $a \in A$ e $x \in U, f(x, a) \in V_a$.

Há várias implementações para o algoritmo do K-Modes, o k-Modes tradicional [8] [9], o k-Modes com inicialização baseado na densidade [7] e k-Prototypes que combina K-Modes e K-Means para ser capaz de agrupar dados numéricos e categóricos [9]. Para este trabalho, foi utilizado o K-Modes com inicialização baseado na densidade que é definida a seguir:

$$d_n(z_i, x_i) = \sum_{a \in A} \phi_a(z_i, x_i), \quad \textbf{(1)}$$

onde

$$\phi_a(z\_i, x\_i) = \begin{cases} 1, & se\ f(z_l, a) \neq f(x_l, x_i), \\ 1 - \frac{|C_{l,a}|}{|C_l|}, & caso\ contrário. \end{cases} \quad \textbf{(2)}$$

Onde $|C_l|$ é o número de objetos no cluster e $|C_{l,a}|$ é o número de objetos com categoria $f(z_l,a)$ do atributo *a* do cluster *l*.

2.3.1 Implementação Python dos algoritmos de agrupamento de k-Modes

PyPI é um repositório de bibliotecas e aplicações desenvolvidos na linguagem de programação Python [10]. A biblioteca Python do kmodes [4] possui as três formas de implementação do algoritmo: a tradicional, a baseada em densidade e a híbrida que envolve dados categóricos e numéricos. O código é modelado baseando nos algoritmos de clustering do scikit-learn [11]. A biblioteca pode ser instalada fazendo uso do sistema repositório de bibliotecas Python chamado PIP, através do comando: pip install kmodes.

2.3.2 Método Elbow para achar o melhor valor do parâmetro K

Este método é uma técnica para encontrar o valor ideal do parâmetro k de um algoritmo K-Means ou K-Modes. O método testa a variância dos dados em relação ao número de clusters. É considerado um valor ideal de k quando o aumento no número de clusters não representa um valor significativo de ganho [12].

2.1 TRABALHOS RELACIONADOS

Data Center é um ambiente crítico em que milhares de operações com dados são processados a todo momento. A fim de buscar melhores formas de se minimizar falhas, diversos estudos são constantemente publicados sobre como melhorar as características de infraestrutura referente a disponibilidade e desempenho operacional dos Data Centers, como é o caso do guia "Data center infrastructure management" [13] e do artigo "High performance datacenter networks: Architectures, algorithms, and opportunities" [14].

Para poder garantir a disponibilidade e desempenho, primeiro precisa-se monitorar. A capacidade e a forma de se monitorar um Data Center é outra área bastante abordada por pesquisadores, como é o caso de pesquisas como "Data Center Workload - Monitoring, Analysis, and Emulation" [15] ou "Performance monitoring of Virtual Machines (VMs) of type I and II hypervisors with SNMPv3" [16].

Como se sabe, o monitoramento de Data Centers pode gerar o registro de diversos incidentes, muito maior que a capacidade de tratamento pelo time de TI. Há um estudo sobre aplicação técnicas de agrupamentos para chamados de TI, "Clustering and labeling IT maintenance tickets" [17] e outro estudo sobre triagem de incidentes em ativos na nuvem, "DeepTriage: Automated Transfer Assistance for Incidents in Cloud Services" [18] que pode ser um bom indicativo do uso de técnicas de agrupamento para a análise de incidentes de Data Center.

3 MATERIAIS E MÉTODOS

Esta seção com suas respectivas subseções refere-se aos materiais, ferramentas e métodos utilizados no decorrer deste projeto, bem como a aplicação da metodologia K-Modes.

3.1 DESCRIÇÃO DA BASE DE DADOS

Os dados foram extraídos a partir da ferramenta de monitoramento Zabbix. Através da interface de exibição de problemas no Zabbix, é possível

exportar uma lista de incidentes em formato de arquivo CSV.

Quadro 1 – Atributos presentes na base de incidentes exportados da ferramenta de monitoramento Zabbix.

ATRIBUTOS	DESCRIÇÃO
Severity	Nível de gravidade do incidente, podendo variar de informação, baixo, médio, alto e desastre.
Time	Momento em que o incidente surgiu.
Recovery Time	Momento em que o incidente foi resolvido.
Status	Atributo que mostra se o incidente continua ativo ou se foi resolvido.
Host	Nome do servidor ou da máquina virtual.
Problem	Nome do problema.
Duration	Tempo total de duração do incidente.
Ack	Atributo que identifica se o incidente foi reconhecido pelos usuários.
Actions	Ações automática ou manuais registrados no sistema para o determinado incidente.
Tags	São etiquetas com informações personalizadas definidas pelo usuário do sistema.

Quadro 2 – Atributos personalizados que podem ser criados pelos administradores do sistema Zabbix e que estavam presentes na base de dados exportada.

ATRIBUTOS	DESCRIÇÃO
AMBIENTE	Informa se o servidor ou máquina virtual é utilizado como um ambiente de desenvolvimento ou de produção.
DEPARTAMENTO	Informa a qual departamento da empresa o servidor ou máquina virtual está associado.
SISTEMA OPERACIONAL	Descreve se o sistema operacional é Windows ou Linux.

Foram exportados registros de um período de três meses consecutivos envolvendo 163 servidores e máquinas virtuais e totalizando 4567 incidentes.

Os dados brutos exportados possuem por padrão 10 atributos conforme o Quadro 1. O atributo Etiqueta é um tipo de dado especial em que o usuário pode definir subatributos personalizados para cada host ou tipo de alarme cadastrado no sistema. Desta forma, o atributo etiqueta na base de incidentes exportada pode ser composto na verdade de vários subatributos a depender da configuração realizada pelo administrador do sistema.

3.1.1 Atributos de Etiqueta

As Etiquetas (Tags) que estão na base de incidentes exportadas do sistema estão definidas conforme o Quadro 2.

3.2 PRÉ-PROCESSAMENTO DOS DADOS

Contanto com as etiquetas definidas pelos usuários do sistema de monitoramento, a base de dados exportada possui 12 atributos. Pode-se dividir os atributos em duas categorias: as que se referem a sua própria natureza e aos que estão relacionados a solução do incidente. Para esse trabalho, o interesse recai nos atributos que descrevem o problema e não como ou quanto tempo durou para ele ser resolvido. Desta forma foram utilizados apenas os atributos:

- Severidade
- Problema
- Ambiente
- Departamento
- Sistema Operacional

É interessante notar que todas as colunas selecionadas possuem apenas dados não numéricos, dessa forma foi mais apropriado o uso da técnica K-Modes ao invés do K-Means ou K-Prototypes.

A coluna Problema possuía uma variedade grande valores pois o nome do problema estava quase sempre composto do nome do host associado. Os dados dessa coluna foram tratados para que o dado representasse uma categoria do problema ao invés do título do problema em si. Dessa forma, foi criada uma classificação para os dados de problema que se resume a 8 tipos:

- Sistema: refere-se à indisponibilidade da aplicação principal valor a organização instalada no host e monitorada pela ferramenta de monitoramento.
- Processador: refere-se a problemas de alto uso e falta de desempenho do processador do host.
- Processos: refere-se aos problemas associados ao desempenho dos

DOI: 10.25286/repa.v7i2.2221

processos que rodam em background no sistema operacional.
- Disco: refere-se a um problema de baixo desempenho e falta de espaço em discos em que os hosts estão hospedados.
- Rede: refere-se apenas aos problemas de latência de rede.
- Host: problemas de indisponibilidade de uma máquina virtual ou servidor físico.
- Backup: problema relativo a um baixo desempenho do procedimento de backup realizado nos hosts.
- Segurança: problemas associados ao atraso na instalação de pacotes de correções e alterações de arquivos de segurança do sistema operacional.

Os dados omissos referentes as colunas Ambientes, Departamento e Sistema Operacional que são atributos do tipo Etiqueta foram tratados e preenchidos com as informações corretas. Para a coluna Departamento, foi realizado um ofuscamento dos dados.

3.3 APLICAÇÃO DA TÉCNICA K-MODES

Para a aplicação da técnica de agrupamento categórico fazendo uso da biblioteca kmodes em Python é necessário a definição dos valores de seis parâmetros que podem influenciar na qualidade final do resultado.

O parâmetro max_iter corresponde ao número máximo de iterações do algoritmo de K-modes para uma execução única. O valor padrão estabelecido na implementação da biblioteca K-Modes em Python é 100. Foi utilizado de forma empírica o valor padrão para este projeto.

O parâmetro cat_dissim corresponde a função de dissimilaridade usada pelo algoritmo de k-modes para variáveis categóricas. Foi adotado a função de dissimilaridade que é o padrão adotado pela biblioteca K-Modes em Python.

O parâmetro init se refere ao método para inicialização do k-modes utilizado. Pode assumir três valore distintos:

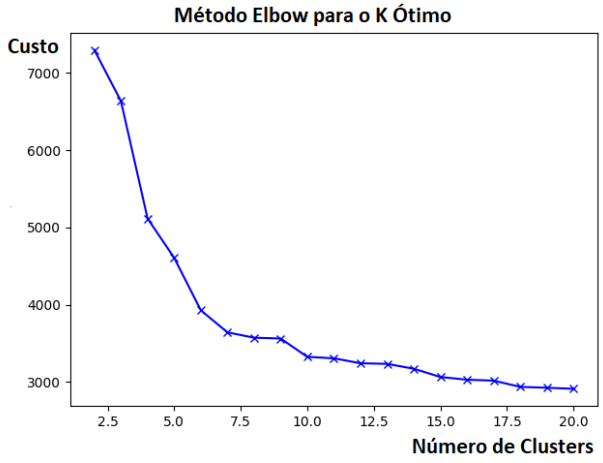
- 'Huang': método do autor Huang publicado em dois trabalhos de 1997 e 1998 [9] [8]. Complexidade de tempo linear. Não adequado em problemas quando os clusters são de tamanhos e densidades diferentes.
- 'Cao': método dos autores Cao, F., Liang, J, Bai, L. publicado em 2009 [7].
- 'random': trabalha com a inicialização aleatória, obtém resultados não repetíveis, não garante agrupamento exclusivo e a escolha inadequada pode resultar em estruturas de cluster altamente indesejáveis.

Huang usa o método de frequência, diferentemente do método Random que seleciona um conjunto arbitrário de pontos do conjunto de dados. Tanto o método de Huang como o Random não consideram as frequências relativas dos atributos no centroide do cluster, produzindo grupos com objetos mais díspares [19]. Os resultados deste trabalho são baseados no método Cao que difere em considerar a densidade de um atributo em um cluster.

O parâmetro n_init corresponde ao número de tempo em que o algoritmo de k-Modes será executado com diferentes sementes de centroide. O valor padrão adotado pela biblioteca K-Modes em Python é 10, o mesmo utilizado para este trabalho.

O parâmetro n_clusters corresponde o número de clusters a serem formados, bem como o número de centroides para gerar. O valor adotado foi obtido pelo método Elbow explicado na seção 3.2.2.

Figura 1 – Método Elbow para achar melhor valor de K



3.2.1 Aplicação do Método Elbow para achar melhor valor de K

Para determinar o parâmetro número ideal de clusters foi adotado o método Elbow modificado para usar o valor diferença dentro do cluster. A partir dos resultados da plotagem das diferenças dentro do cluster para vários valores, o princípio do método Elbow obtém o valor de k no ponto em que

o valor não diminui significativamente com a adição do valor de k.

De acordo com a Figura 1, é possível perceber que o valor 10 é um número adequado de quantidade de clusters k. Desta forma, foi considerado escolher k = 10 para a análise para agrupamento dos dados. Isso significa que serão formados dez grupos de incidentes com base nas características dos hosts.

4 ANÁLISE E DISCURSÃO DOS RESULTADOS

Nesta seção são apresentados os resultados e a análise realizada pelos autores deste trabalho e pelo especialista em NOC da organização em que a base foi extraída.

4.1 RESULTADOS

Após o procedimento de mineração de dados realizado, foi elaborada uma Tabela de Correlação entre os atributos para cada clusters gerado. Esse tipo de correlação auxilia na extração de possíveis regras e informações não triviais associadas aos incidentes de servidores e máquinas virtuais.

Tabela 1- Clusters x Atributos: distribuição dos incidentes entre os clusters.

#	AMB	DEPT	PROB	SEV	S.O.	QTDE
0	PROD	D5	HOST	ALTA	WIN	1354
1	PROD	D6	DISCO	BAIXA	LINUX	764
2	PROD	D6	DISCO	MÉDIA	WIN	457
3	DEV	D5	SISTEMA	ALTA	LINUX	978
4	PROD	D6	REDE	BAIXA	WIN	274
5	PROD	D7	SISTEMA	ALTA	WIN	407
6	DEV	D4	DISCO	ALTA	WIN	126
7	PROD	D6	DISCO	ALTA	LINUX	58
8	PROD	D5	DISCO	ALTA	LINUX	11
9	PROD	D2	DISCO	ALTA	WIN	138

A Tabela 1 é um resumo da distribuição dos incidentes em que é comparado todos os atributos com maior valor de frequência para cada cluster. Ao analisar a Tabela, foi possível perceber que:

- Os problemas relacionados a disco estão presentes nos hosts associados a maioria dos departamentos.
- Os problemas de indisponibilidade de host e de aplicações hospedadas nos hosts impactam principalmente nos departamentos 5 e 7.
- Os problemas relacionados a desempenho de rede parecem afetar de forma majoritária os ambientes de produção com sistema operacional Windows.

A Tabela de Correlação foi criada a partir da síntese das informações extraídas dos dados que também estão representados pelos gráficos 2 ao 6. Cada gráfico exibe a distribuição dos incidentes (eixo das ordenadas) por um atributo individual para cada cluster (eixo das abcissas). Apesar da análise correlacionada dos atributos ser maior valor, avaliar a distribuição de cada atributo pelos agrupamentos mostra-se também útil para extração de informações relevantes.

As Figuras 2 a 6 foram disponibilizadas no apêndice deste trabalho para melhor visualização. Na Figura 2, é possível avaliar a distribuição dos incidentes quanto ao atributo severidade. Os incidentes de severidade alta possuem uma grande representatividade principalmente nos clusters 0, 3 e 5. Os incidentes do tipo atenção estão presentes mais nos clusters 1 e 4. Os incidentes de severidade média estão mais presentes no cluster 2. Na Figura 3 é possível analisar a distribuição dos incidentes sob a perspectiva do atributo problema. Há várias informações relevantes no gráfico, por exemplo, o cluster 5, há eventos de indisponibilidade do sistema instalado no host, mas não sobre a indisponibilidade do host em si. Já no cluster 3, uma parte relevante da indisponibilidade do sistema pode ter sido causada pela própria indisponibilidade do host. Na Figura 4, há a distribuição dos incidentes referente ao atributo Ambiente. É importante notar que a quantidade de host do tipo produção é muito superior que ao do tipo desenvolvimento. No cluster 3, pode-se dizer que o subconjunto de incidentes afetou tanto o ambiente de desenvolvimento quanto de produção. Já no cluster 6, temos os incidentes que só afetaram o ambiente de desenvolvimento. A Figura 5 se refere ao atributo departamento. Uma forma de analisar esse gráfico é verificar quais subconjuntos de incidentes que afetaram a maioria dos departamentos e quais afetaram exclusivamente um departamento específico. Por fim, na Figura 6, há a distribuição dos incidentes quanto ao atributo sistema operacional. Ao analisar os clusters, pode-se dizer que, em geral, os incidentes que afetam os hosts do Windows são distintos dos incidentes que afetam os hosts com Linux.

DOI: 10.25286/repa.v7i2.2221

4.1 PLANO DE AÇÃO

A partir dos resultados analisados foi possível encontrar alguns indicativos de problemas que estão afetando os hosts de uma maneira geral. Através do auxílio de um especialista no Data Center da organização em que foi extraída a base de dados para analisar a Tabela de Correlação, foi também possível criar um plano de ação como sugestão para tomada de decisão pela gerência do departamento de TI da empresa. O plano de ação consiste nos seguintes passos:

1. Avaliar a possibilidade de se investir em uma solução de equipamento de *storage all-flash* a ser adicionado ao ambiente de Data Center *on-premise* ou migrar alguns serviços e sistemas para nuvem.
2. Investigar o motivo pelo qual há uma predominância de eventos de desempenho de rede no ambiente Windows.
3. Investigar por que as máquinas relacionadas ao DEPT05 possuem tantos alarmes de host.

O primeiro item do plano de ação pode ser categorizado como uma ação de investimento em infraestrutura e os custos desse investimento podem ser, em parte, justificados através dos resultados obtidos deste trabalho. Os itens 2 e 3 são ações que apesar de ainda não serem a causa-raiz dos incidentes, irão servir para evidenciar a origem dos problemas.

4.2 DISCUSSÃO

O uso de métodos de agrupamentos de dados para encontrar informações relevantes sobre incidentes de Data Center parece uma estratégia promissora para uso em tomadas de decisão para gestão e investimentos em infraestrutura TI.

Pela natureza dinâmica dos eventos de incidentes de Data Center, os resultados finais dependem do recorte do período de eventos extraídos da ferramenta de monitoramento. O plano de ação gerado poderia ter sido mais aprofundado se a base de dados analisada fosse composta também por outros tipos de hosts como switches, links de Internet, storages e outros tipos de equipamentos comumente presentes em um ambiente de Data Center.

É também um indicativo que criar uma etiqueta chamada Categoria de Problema irá auxiliar no pré-processamento dos dados de forma mais eficiente. Da mesma forma, parece ser relevante para avaliação de desempenho que a base de dados possua os atributos de quantidade de CPU, memória e disco associados aos servidores e máquinas virtuais para cada incidente registrado.

Para este trabalho, foi utilizada a base de dados de incidentes da ferramenta Zabbix, mas parece uma estratégia indicada a aplicação de métodos de agrupamento (k-means, k-modes e k-prototypes) para base de incidentes de qualquer natureza, não se limitando a área de Data Center.

A biblioteca kmodes utilizada neste trabalho possui a capacidade de se processar os dados de forma paralela. Isso pode definido através dos parâmetros de configuração. Apesar dessa funcionalidade não ter sido utilizada para este trabalho, pois a base utilizada para a análise é pequena e o tempo de execução algoritmo é curto, essa funcionalidade mostra-se relevante para viabilizar análise em big data.

5 CONCLUSÕES

O principal objetivo deste trabalho foi aplicar a técnica de mineração de dados chamada K-Modes a fim de achar evidências das causas-raiz dos incidentes dos três últimos meses referente aos servidores e máquinas virtuais de uma organização.

Neste trabalho foi possível agrupar os incidentes mais relevantes classificando-os por características semelhantes. O método utilizado demonstrou-se uma técnica adequada no agrupamento de incidentes com valores de atributos semelhantes entre si e na criação de uma Tabela de Correlação dos tipos de incidentes por atributos para os hosts do Data Center.

Como trabalho futuro, é um bom indicativo analisar os atributos que se referem a forma como os incidentes são resolvidos pois poderão revelar informações importantes sobre a eficiência e gargalos no ambiente NOC da organização. É provável que o método k-Prototypes seja mais adequado para esse cenário pela capacidade de suportar dados categóricos e numéricos ao mesmo tempo. Outro ponto que irá agregar mais valor aos resultados é ampliar o escopo da base utilizada para envolver outros tipos de hosts na análise, pois as causas-raiz dos problemas podem ter origem nos hosts não analisados anteriormente.

REFERÊNCIAS

[1] MILOSLAVSKAYA, N. Network security intelligence center as a combination of SIC and NOC. Procedia Computer Science, v. 145, p. 354-358, 2018.

[2] ZABBIX LLC. Zabbix features overview: What is Zabbix. [S. l.], 2021. Disponivel em: <https://www.zabbix.com/features>. Acesso em: 29 July 2021.

[3] SERPRO. Zabbix: solução livre para gestão de serviços e sistemas. SERPRO, 2011. Disponivel em: <https://www.serpro.gov.br/menu/noticias/noticias-antigas/zabbix-solucao-livre-para-gestao-de-servicos-e-sistemas>. Acesso em: 9 Agosto 2021.

[4] DE VOS, N. J. kmodes categorical clustering library, 2015. Disponivel em: <https://github.com/nicodv/kmodes>. Acesso em: 09 August 2021.

[5] LIANG, J. Y.; LI, D. Y. Uncertainty and knowledge acquisition in information systems. **Science Press**, Beijing, 2005.

[6] PAWLAK, Z. **Rough sets:** Theoretical aspects of reasoning about data. [S.l.]: Springer Science & Business Media, 1991.

[7] FUYUAN, C.; LIANG, J.; BAI, L. A new initialization method for categorical data clustering. **Expert Systems with Application**, v. 36, n. 7, p. 10223-10228, 2009.

[8] HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical value. **Data mining and knowledge discovery**, v. 2, n. 3, p. 283-304, 1998.

[9] HUANG, Z. Clustering large data sets with mixed numeric and categorical values. **Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)**, 1997. 21-34.

[10] MILLMAN, K. J.; AIVAZIS, M. Python for scientists and engineers. **Computing in Science & Engineering**, v. 13, n. 2, p. 9-12, 2011.

[11] PEDREGOSA, F. E. A. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, v. 2, p. 2825-2830, 2011.

[12] BHOLOWALIA, P.; KUMAR, A. EBK-means: A clustering technique based on elbow method and k-means in WS. **International Journal of Computer Applications**, v. 105, n. 9, 2014.

[13] COLE, D. **Data center infrastructure management**. Data Center Knowledge. [S.l.]. 2012.**of Computer Applications**, v. 105, n. 9, 2014.

[14] ABTS, D.; KIM, J. High performance datacenter networks: Architectures, algorithms, and opportunities. **Synthesis Lectures on Computer Architecture**, v. 6, n. 1, p. 1-115, 2011.

[15] MOORE, J. E. A. **Data center workload monitoring, analysis, and emulation**. Eighth Workshop on Computer Architecture Evaluation using Commercial Workloads. [S.l.]: [s.n.]. 2005. p. 1-8.

[16] IQBAL, A.; PATTINSON, C.; KOR, A.-L. **Performance monitoring of Virtual Machines (VMs) of type I and II hypervisors with SNMPv3**. World Congress on Sustainable Technologies (WCST). [S.l.]: IEEE. 2015. p. 98-99.

[17] ROY, S. et al. Clustering and labeling IT maintenance tickets. **International Conference on Service-Oriented Computing**, 2016. 829-845.

[18] PHAM, P. et al. DeepTriage: Automated Transfer Assistance for Incidents in Cloud Services. **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, 2020. 3281-3289.

[19] LLOP TORRENT, N. The K-modes algorithm applied to Gender Analysis, 2019.

[20] DONADIO, P.; CIMMINO, A.; PRASAD, R. A cloud infrastructure to manage future internet: The virtual network operation center. **Journal of Green Engineering**, 15 March 2011. 255-265.

[21] ZABBIX LLC. 2 Problems: Zabbix Documentation 5.0. [S. l.], 2021. Disponivel em: <https://www.zabbix.com/documentation/5.0/manual/web_interface/frontend_sections/monitoring/problems>. Acesso em: 29 July 2021.

[22] HAN, J.; PEI, J.; KAMBER, M. **Data mining:** concepts and techniques. 3ª. ed. [S.l.]: Elsevier, 2011.

DOI: 10.25286/repa.v7i2.2221

[23] BARROSO, L. A.; HÖLZLE, U. The datacenter as a computer: An introduction to the design of warehouse-scale machines. **Synthesis lectures on computer architecture**, v. 4, n. 1, p. 1-108, 2009.

APÊNDICE

Figuras com as análises para cada atributo.

Figura 2 – Gráfico quantidade de incidentes x clusters x severidade

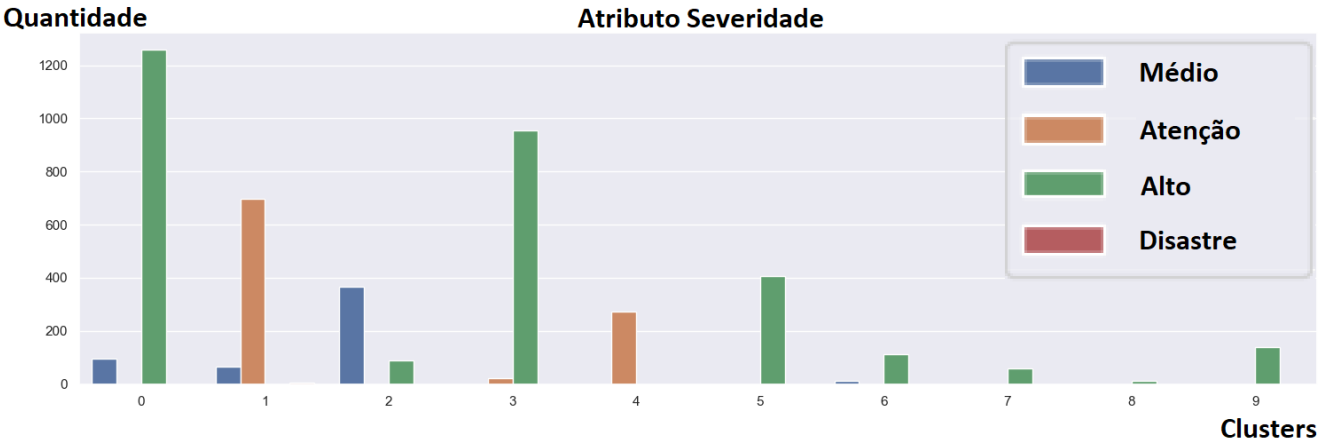


Figura 3 – Gráfico quantidade de incidentes x clusters x categoria de problema

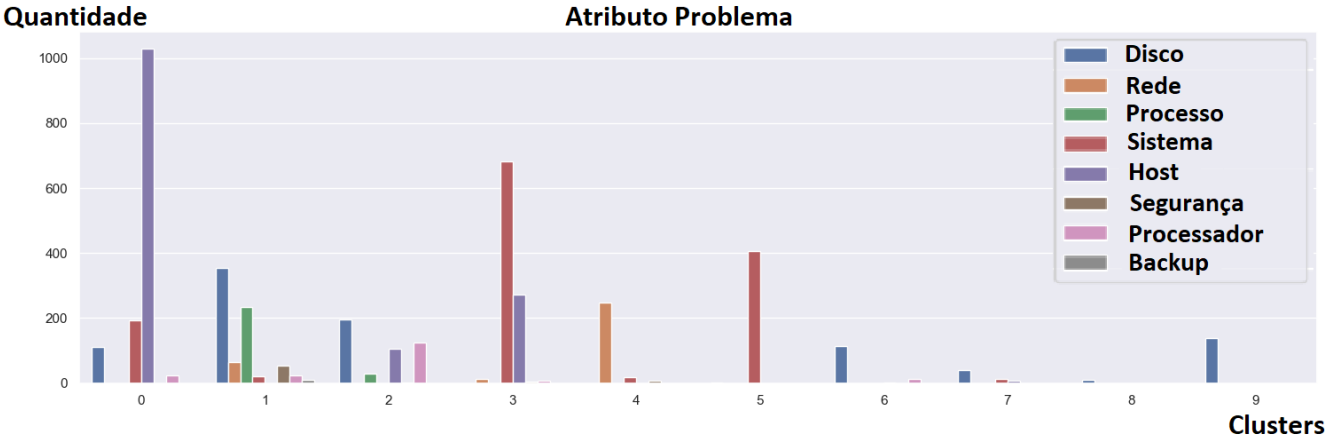


Figura 4 – Gráfico quantidade de incidentes x clusters x ambiente

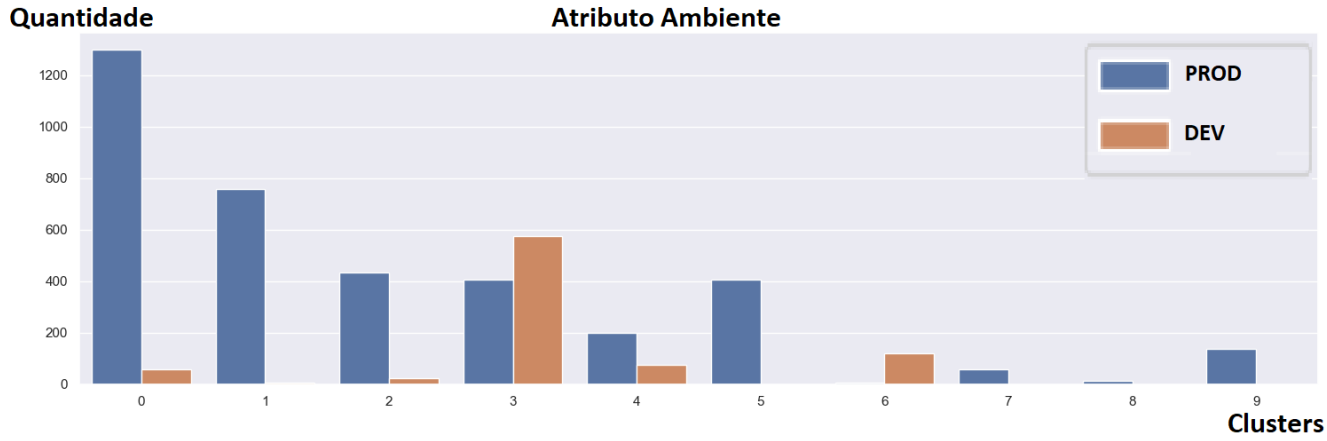


Figura 5 – Gráfico quantidade de incidentes x clusters x departamento

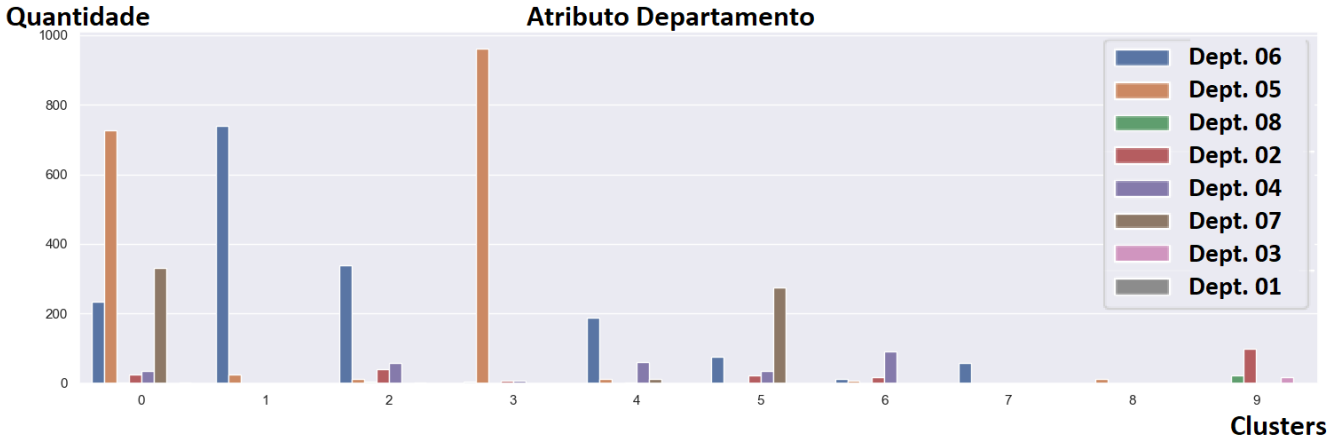
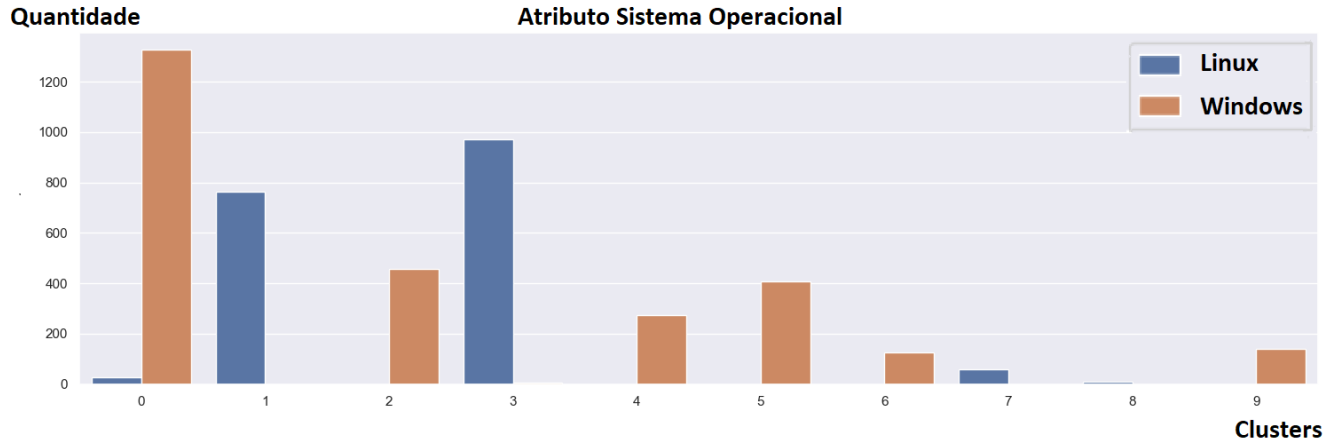


Figura 6 – Gráfico quantidade de incidentes x clusters x sistema operacional



DOI: 10.25286/repa.v7i2.2221