



Datamining para Auditoria de Segurança

António Gonçalves

Fonte: Slides do Prof. Victor Lobo

Mestrado em Segurança da Informação e Direito no Ciberespaço



Datamining para Auditoria de Segurança

1. Detecção de Anomalias e Comportamentos Suspeitos

- **Problema:** É difícil identificar padrões anómalos em grandes volumes de dados de registos (logs), especialmente em tempo real.
- **Desafio:** Separar falsos positivos de verdadeiras ameaças.
- **Objetivo:** Desenvolver algoritmos de mineração de dados para reconhecer comportamentos fora do padrão que possam indicar intrusões ou acessos não autorizados.

2. Identificação de Ameaças Internas

- **Problema:** As ameaças internas (insiders) são difíceis de detetar, uma vez que os utilizadores internos já têm permissões.
- **Desafio:** Analisar padrões comportamentais e identificar desvios no uso habitual dos sistemas.
- **Objetivo:** Usar técnicas de clustering e análise preditiva para identificar perfis de risco entre os utilizadores internos.



Datamining para Auditoria de Segurança

3. Correlação de Eventos de Segurança

- **Problema:** A correlação manual de eventos de segurança em grandes volumes de logs é demorada e ineficaz.
- **Desafio:** Agregar e analisar eventos provenientes de diferentes fontes (firewalls, IDS, sistemas operativos) para identificar possíveis ataques coordenados.
- **Objetivo:** Utilizar técnicas de associação e descoberta de padrões sequenciais para correlacionar eventos.

4. Prevenção de Fraudes

- **Problema:** As fraudes internas ou externas são muitas vezes descobertas tarde demais.
- **Desafio:** Identificar padrões comportamentais que indiquem tentativas de fraude.
- **Objetivo:** Implementar algoritmos de classificação supervisionada (como Decision Trees e Random Forest) para detetar transações suspeitas.



Datamining para Auditoria de Segurança

5. Redução de Falsos Positivos e Falsos Negativos

- **Problema:** Sistemas de auditoria tradicionais geram muitos alertas irrelevantes (falsos positivos) ou não detetam comportamentos perigosos (falsos negativos).
- **Desafio:** Melhorar a precisão dos modelos de detecção de intrusões (IDS) usando técnicas avançadas de mineração de dados.
- **Objetivo:** Treinar modelos com conjuntos de dados balanceados e aplicar técnicas como SVM (Support Vector Machines) e redes neuronais.

6. Análise de Riscos e Vulnerabilidades

- **Problema:** É difícil prever quais vulnerabilidades podem ser exploradas.
- **Desafio:** Priorizar as vulnerabilidades que representam maior risco.
- **Objetivo:** Utilizar técnicas de clustering e scoring para classificar vulnerabilidades com base no seu potencial impacto.

Problema:

- Como detectar intrusões, quando não sabemos o que são ? Quando não temos “assinaturas” ? (3º passo no framework NIST)
- Caso 1: conhecemos casos passados em que foram detectadas intrusões, mas há pequenas variações...
- Caso 2: conhecemos muitos casos “normais”, que variam muito entre si, mas não sabemos o que poderá acontecer de diferente
 - Ficheiros ou ligações normais/anormais
 - Padrões de tráfego normais/anormais



Os 4 Passos do Data Mining para Auditoria de Segurança em Sistemas:

1. Recolha e Pré-processamento de Dados:

- Recolher logs e registos de segurança.
- Limpar, integrar, transformar e anonimizar os dados.

2. Exploração e Análise de Padrões:

- Aplicar algoritmos de classificação, clustering, regras de associação e deteção de anomalias.

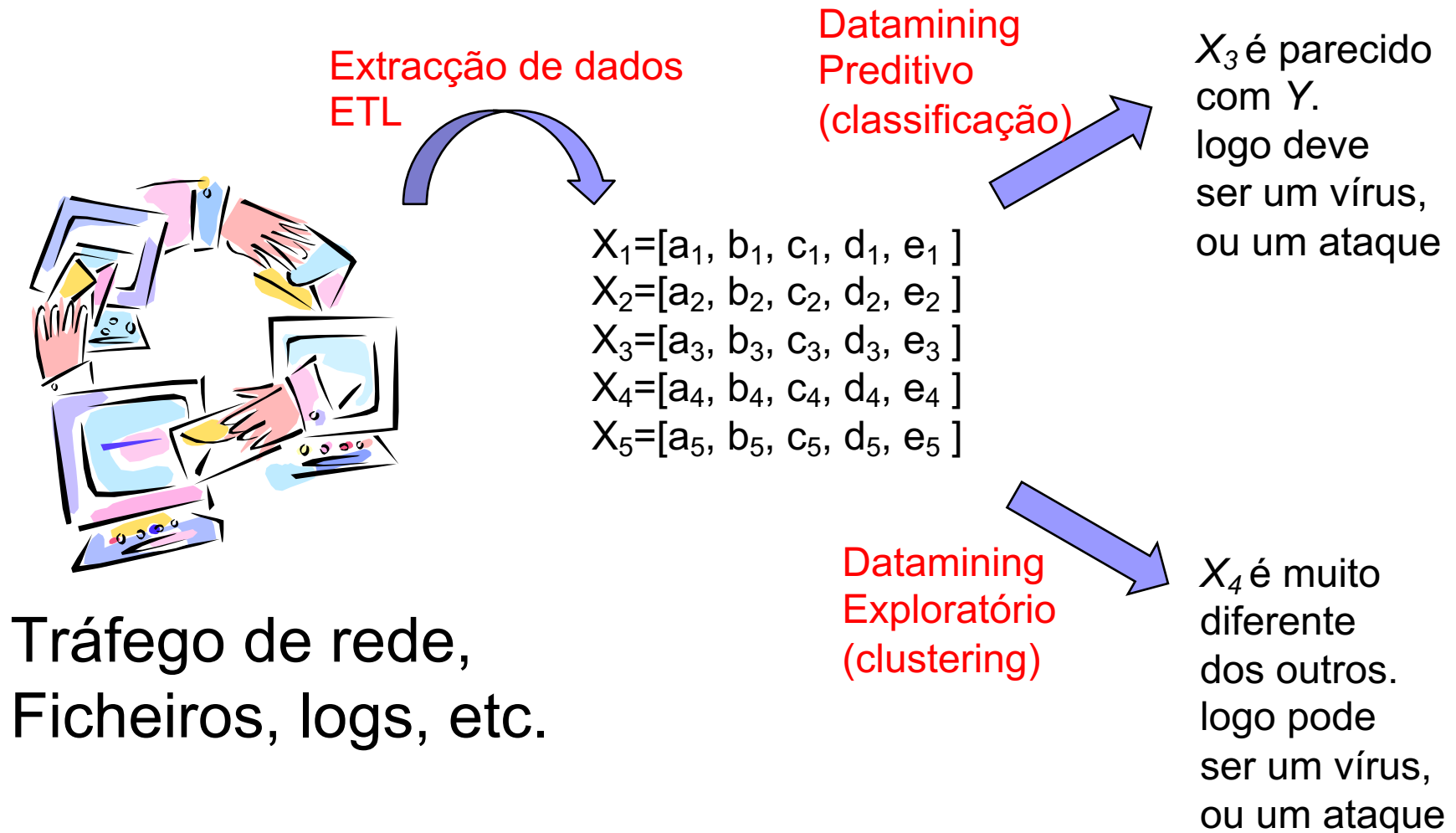
3. Interpretação e Avaliação de Resultados:


- Avaliar modelos com métricas (precisão, recall, F1-score) e analisar falsos positivos/negativos.

4. Implementação e Monitorização Contínua:

- Implementar os modelos nas auditorias e ajustar com base nos novos dados e ameaças emergentes.

Ideia geral





Programa (traços gerais)

- Introdução às técnicas para deteção e classificação de cyber-ameaças usando datamining (parte inicial)
- Introdução ao **datamining** e **pré-processamento** de dados
- Técnicas de **visualização** de dados **multi-dimensionais**
- Técnicas de **deteção de outliers** e **comportamentos anormais**
- Técnicas de **classificação** de comportamentos
- Técnicas para deteção e classificação de cyber-ameaças (parte final)





Método de Avaliação

- “Repetição escrita”
 - 45% da nota
- Apresentação oral e resumo de um artigo
 - 30% da nota
- Projecto de DM para Auditoria de segurança
 - 25% da nota

Método de Avaliação - Datas

EVENTO	DATA	DIA SEMANA
Submissão Artigo	21/03/2025	Sexta-Feira
Apresentação Artigo	24/03/2025	Segunda-Feira
Proposta Projeto	02/05/2025	Sexta-Feira
Submissão Projeto	23/05/2025	Sexta-Feira
Defesa Projeto	26/05/2025	Segunda-Feira
Repetição Escrita	02/06/2025	Segunda-Feira

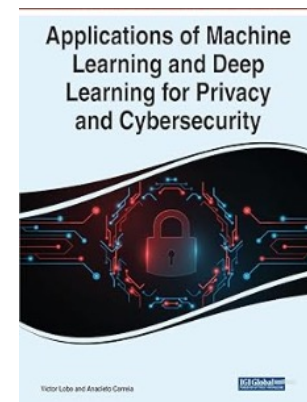
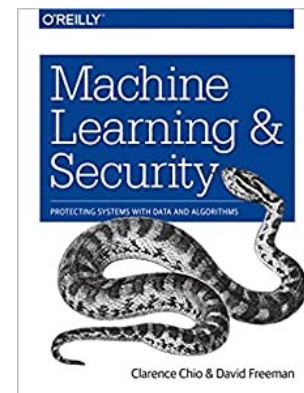


Informação Variada

- Aulas a distância (3ºf das 18:00 às 20:00)
 - [Zoom](#)
- Dúvidas:
 - agoncalves@tecnico.ulisboa.pt
- Apoio
 - Depois das aulas
 - Por email
- Material de apoio (GitHub)
 - [Link](#)

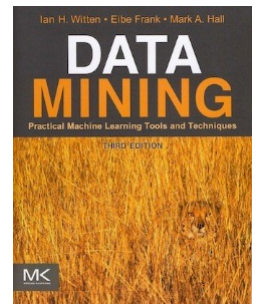
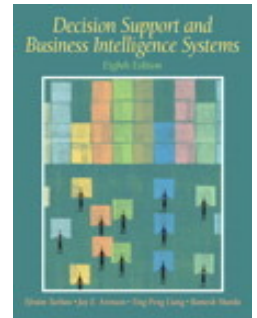
Bibliografia

- Livros de textos (não são seguidos “à risca”)
 - Textos de apoio disponíveis no site da UC
 - **Machine Learning and Security: Protecting Systems with Data and Algorithms**, Clarence Chio, David Freeman, O'Reilly Media, 2018
cap.1,2,3,5
 - **Hands-On Machine Learning for Cybersecurity**; Soma Halder, Sinan Ozdemir, Packt Publishing, 2018
 - **Applications of Machine Learning and Deep Learning for Privacy and Cybersecurity**, Victor Lobo, Cortez e Correia, IGI Global, 2022



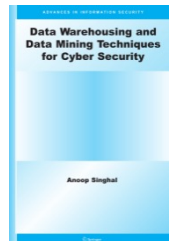
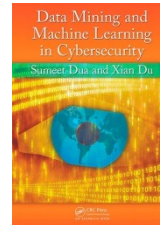
Bibliografia

- **Decision Support and Business Intelligence Systems**, Turban, E., J. E. Aronson, et al., Prentice Hall, 2010
- **Data mining: practical machine learning tools and techniques**; Ian H. Witten, Eibe Frank, Mark A. Hall: Morgan Kaufmann, 2011 ([WEKA](#))
- **Python Machine Learning**: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, Raschka, Packt Pub., 2019



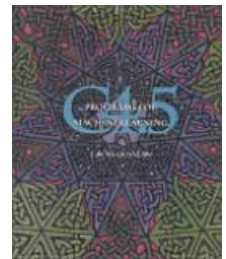
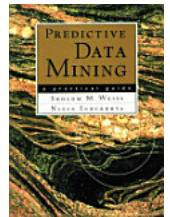
Bibliografia mais especializada

- **Data Mining and Machine Learning in Cybersecurity**, Sumeet Dua, Xian Du, ISBN: 978-1439839423, Auerbach Publications, 2011
- **Data Mining Tools for Malware Detection**, Mehedy Masud, Latifur Khan, Bhavani Thuraisingham , ISBN: 978-1439854549, Auerbach Publications 2011.
- **Data Warehousing and Data Mining Techniques for Cyber Security**, Anoop Singhal, ISBN: 978-0387264097, Springer 2006.
- **Applications of Data Mining in Computer Security**, Barbará, Daniel; Jajodia, Sushil (Eds.), ISBN: 978-1-4020-7054-9, Springer 2002.



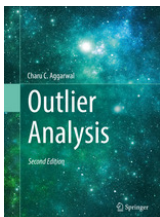
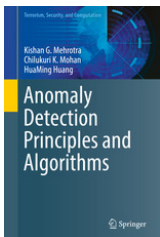
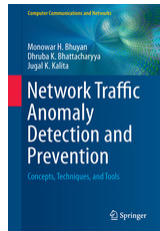
Bibliografia geral de DM

- **Machine Learning**, Tom M. Mitchell, McGraw Hill, 1997
- **Pattern Classification**, Duda, Hart, & Stork, Wiley, 2001
- **Principles of data mining**, David. J. Hand, Heikki Mannila, Padhric Smyth, MIT Press, 2001
- **Predictive data mining**, Sholom M. Weiss, Nitin Indurkha, Morgan Kaufmann, 1997
- **C4.5: Programs for Machine Learning**, John Ross Quinlan, Morgan Kaufmann, 1992



Bibliografia

- Network Traffic Anomaly Detection and Prevention - Concepts, Techniques, and Tools, Bhuyan, Monowar H., Bhattacharyya, Dhruba K., Kalita, Jugal K., ISBN: 978-3-319-65188-0, Springer 2017
- Anomaly Detection Principles and Algorithms, Mehrotra, Kishan G., Mohan, Chilukuri, Huang, Huaming, 978-3-319-67526-8, Springer 2017
- Outlier Analysis, Aggarwal, Charu C., 978-3-319-47578-3, Springer 2017
- Network Intrusion Detection and Prevention - Concepts and Techniques, Ghorbani, Ali A., Lu, Wei, Tavallaei, Mahbod, 978-0-387-88771-5, Springer 2010



Outros sites interessantes...

- Decisionarium
 - Software GNU, referências, etc
 - <http://www.decisionarium.tkk.fi>
- DSS Resources
 - Prof. Daniel Power, livros, referências, etc
 - <http://dssresources.com/>
- Machine Learning Network
 - www.mlnet.org
 - Software, dados, conferências, projectos, etc.
- Fabricantes de soluções “dedicadas”
 - Para gestão de terrenos, para marketing, etc, etc

Repositórios de dados

■ Repositório de Irvine (UCI)

- <https://archive.ics.uci.edu/ml/index.php>
- Dados, software, artigos
- Um clássico! Um “must” !



■ Repositório Kaggle

- www.kaggle.com/datasets
- Muito actual, muito activo



■ Repositório do IEEE

- IEEE Data Port
- <https://ieee-dataport.org/datasets>



■ Repositório para Cibersegurança

- ICSX: <http://www.iscx.ca/datasets/> (mas o KDD99 está disponível no UCI)

Resolução de problemas práticos

■ MS-Excel

- Todos conhecem !
- Resolve a maioria dos problemas simples

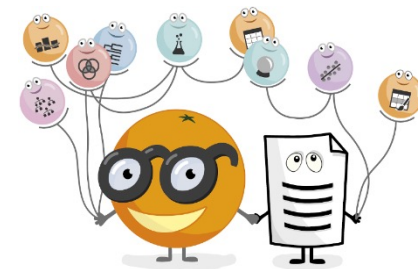
■ WEKA

- Java, free, <https://www.cs.waikato.ac.nz/ml/weka/>
- Muitos algoritmos, bem documentados



■ Orange

- Python, free, <https://orange.biolab.si>
- Interface gráfica



■ Outros

- MATLAB, R, Skikit-learn, Keras.SPSS e Clementine, SAS Enterprise Miner, IBM Intelligent Miner, SAP BI...,

Resolução de problemas práticos

- **Google colab:** plataforma baseada em nuvem que permite a execução de código Python diretamente a partir de um navegador. É especialmente útil para projetos
- 1. **Pandas:** Manipulação e análise de dados (DataFrames, EDA).
- 2. **NumPy:** Cálculos numéricos e manipulação de arrays.
- 3. **Matplotlib:** Visualizações básicas (gráficos, histogramas).
- 4. **Seaborn:** Visualizações avançadas (matrizes de calor, boxplots).
- 5. **Scikit-learn:** Algoritmos de Machine Learning (classificação, regressão, clustering).
- 6. **XGBoost/LightGBM:** Modelos avançados (boosting, classificação eficiente).
- 7. **Statsmodels:** Análises estatísticas (testes, regressão).
- 8. **TensorFlow/Keras:** Redes neurais e Deep Learning.

Artigos a apresentar

(exemplos... mas **procurem** !)

- Bollmann, C. A., Tummala, M., & McEachen, J. C. (2021). Resilient real-time network anomaly detection using novel non-parametric statistical tests. *Computers & Security, 102*, 102146.
doi:https://doi.org/10.1016/j.cose.2020.10214
- Gibert, D., Mateu, C., Planes, J., & Marques-Silva, J. (2021). Auditing static machine learning anti-Malware tools against metamorphic attacks. *Computers & Security, 102*, 102159.
doi:https://doi.org/10.1016/j.cose.2020.102159
- Krumay, B., Bernroider, E. W. N., & Walser, R. (2018). *Evaluation of Cybersecurity Management Controls and Metrics of Critical Infrastructures: A Literature Review Considering the NIST Cybersecurity Framework*, Cham.
- Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-Based Systems, 78*, 13-21.
doi:https://doi.org/10.1016/j.knosys.2015.01.009

Artigos a apresentar

(exemplos... mas **procurem** !)

- Mitchell, R., & Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Comput. Surv.*, 46(4), Article 55.
doi:10.1145/2542049
- Casas, P., Mazel, J., & Owezarski, P. (2012). Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. *Computer Communications*, 35(7), 772-783.
doi:https://doi.org/10.1016/j.comcom.2012.01.016
- García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, 28(1), 18-18-28.
doi:10.1016/j.cose.2008.08.003

Artigos a apresentar (exemplos...)

- Data Mining for Cyber Security, V.Chandois *et al.*, in Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006.
- Data mining methods for anomaly detection KDD-2005 workshop report, Margineantu *et al.*, ACM SIGKDD Explorations Newsletter, Volume 7 Issue 2, December 2005.
- On the efficacy of data mining for security applications, Ted E. Senator, ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics -CSI-KDD '09, 2009.
- Metrics for mitigating cybersecurity threats to networks, IEEE Internet Computing, 14, 1, Jan-Feb 2010.
- A Combined Fusion and Data Mining Framework for the Detection of Botnets, Kiayias *et al.*, Conference For Homeland Security, 2009. CATCH '09. Cybersecurity Applications & Technology, March 2009
- A study of Spam Detection Algorithms on Social Media Networks, Jacob Soman Saini, International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013.

Artigos a apresentar (...exemplos...)

- Comparative Study of Two- and Multi-Class-Classification-Based Detection of Malicious Executables Using Soft Computing Techniques on Exhaustive Feature Set. Shina Sheen, R. Karthik and R. Anitha; International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013
- Botnets: A Study and Analysis, G. Kirubavathi and R. Anitha, International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013
- The VoIP intrusion detection through a LVQ-based neural network, Zheng Lu ; Taoxin Peng, International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Detection of applications within encrypted tunnels using packet size distributions, Mujtaba, G., Parish, D.J., International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Email classification: Solution with back propagation technique, Ayodele et al. International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Malware detection using statistical analysis of byte-level file content, Tabish et al., CSI-KDD '09 Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics, 2009