ICAI 2023
12th International Conference on Applied Informatics
Eger, Hungary, 2023

# Using data mining methods for deeper analysis of vulnerability testing data

**Ferenc Koczka**[a], **Tibor Tajti**[a]

[a]Eszterházy Károly Catholic University
koczka.ferenc@uni-eszterhazy.hu
tajti.tibor@uni-eszterhazy.hu

## Abstract

Detecting and measuring vulnerabilities in an organization's IT system is essential to developing a well-functioning protection strategy. In larger organizations, it is almost impossible to provide an immediate and correct response to every vulnerability, and it is also impossible to implement all available patches on all elements of the system immediately. Therefore, IT management must develop a strategy to respond to this type of emergency within a reasonable timeframe.

In managing vulnerabilities, cases that are likely to cause a more serious emergency should be given higher priority. A prerequisite for this process is a correct assessment of the severity of the risk posed by the existence of each vulnerability, which requires their comparability and thus the definition of an exact metric. The de facto procedure developed and widely used has been incorporated into several software tools.

A collaborative program to record vulnerabilities and describe their characteristics was launched by the Massachusetts Institute of Technology Research and Engineering (MITRE), resulting in the creation of the Common Vulnerabilities and Exposures (CVE) database. [5] These register vulnerabilities that are recognized or documented by manufacturers and that can be patched independently of other vulnerabilities. Once known, vulnerabilities are provided with a standardized identifier, a short description, and a field for recording comments. However, it does not include technical information on vulnerabilities, nor information on their potential impact or how to fix them. This information, where it exists, can be found partly in lists maintained by the vendors and partly in other databases, such as the US National Vulnerability Database (NVD) [1] or the CERT/CC Vulnerability Notes

database. [3]

The methodologies, supporting standards, and best practices used in risk analysis often do not provide quantitative tools for estimating the risk of individual vulnerabilities. An inescapable contributor to vulnerability metrics is the Common Vulnerability Scoring System (CVSS) framework owned and managed by the Forum of Incident Response and Security Teams, Inc. The primary purpose of CVSS is to quantify the severity of the vulnerability . The set of factors that determine the severity of a vulnerability are grouped by the CVSS into three main metrics, which are categorized as Base, Temporal and Environmental. The Base score remains unchanged throughout the vulnerability lifecycle, the Temporal scores are a function of time, and the Environmental scores correct weight the final score based on the severity of the consequences of an incident on the system.

In the context of an ongoing research project at our university, we are processing the test results of vulnerability testing software, in which we are investigating the applicability of data mining methods. Our dataset contains time-series vulnerabilities of university IT infrastructure, which separately treat system elements open to the public Internet and internal network elements segmented by VLAN settings.

The best approach for data mining in a complex university network depends on the specific goals and characteristics of the data. Combining multiple methods and evaluating their performance may be necessary to find the most effective solution for your data.

Using the aforementioned methodology, we collected information about 745 nodes in the university network. In our research, we conducted data mining using two approaches, cluster analysis, and using dimension reduction algorithms to visualize the multidimensional nodes in 2D.

For cluster analysis, we have chosen the K-Means, the Hierarchical clustering, and the DBSCAN algorithms. [2] K-means is a centroid-based clustering algorithm that groups data points into k clusters based on their distance to the cluster centroid. It is a simple and efficient algorithm that is easy to implement and understand. It is particularly well-suited to clustering data that are well-separated and has spherical clusters of similar size. K-means is sensitive to the algorithm's initial conditions and the number of clusters, and it might not work well when the clusters have different shapes or sizes. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups data points into clusters based on their density. It does not require the number of clusters to be specified and it can find clusters of different shapes. DBSCAN is particularly well-suited to clustering data that has a complex structure, such as clusters of different shapes and sizes, or data with a lot of noise. However, DBSCAN can be sensitive to the density and distance parameters, so it might be necessary to fine-tune the parameters to get the best results. We have used cluster analysis and 2D visualization [4] to extract insights from data and interpret and understand the results of these analyses in the context of the problem and the goals of the analysis. Then we conclude from the results of these analyses. First we identify patterns, by

looking for patterns in the data, such as clusters of similar data points or outliers, and try to understand what these patterns might mean in the context of the problem. Then we compare clusters: examining the characteristics of different clusters, such as the size, shape, and distribution of the data points, and try to understand how these clusters differ from each other. [6] For investigating outliers we look for data points that are not well-represented by the clusters and try to understand why these points might be outliers. Then we try understanding the structure of the data, look for structure in the data, such as clusters, patterns, or outliers, and try to understand how this structure might relate to the problem or the goals of the analysis. 2D visualization can help to understand the distribution of the data in the two selected dimensions and the structure of the data. Finally, combining the results of the cluster analysis and the 2D visualization provide a more complete understanding of the data. The cluster assignments can be used to colour the data points in the 2D visualization, and the results of the 2D visualization can be used to interpret the clusters obtained by the cluster analysis.

# References

[1] H. Booth, D. Rike, G. Witte: *The National Vulnerability Database (NVD): Overview*, en, ITL Bulletin, National Institute of Standards and Technology (2013), Accessed January 29, 2023, URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=915172.

[2] M. Ester, H.-P. Kriegel, J. Sander, X. Xu: *A density-based algorithm for discovering clusters in large spatial databases with noise*, en, KDD 96.34 (1996), pp. 226–231.

[3] S. Houmb, V. Franqueira, E. Engum: *Quantifying security risk level from CVSS estimates of frequency and impact*, en, Journal of Systems and Software 83.9 (2010), pp. 1622–1634.

[4] Maaten, Laurens, G. Hinton: *Visualizing data using t-SNE*, en, Journal of machine learning research 9.11 (2008).

[5] J. Wang, M. Guo, H. Wang, L. Zhou: *Measuring and ranking attacks based on vulnerability analysis*, en, Information Systems and e-Business Management 10.4 (2011), pp. 455–490.

[6] Y. Zhao, G. Karypis: *Evaluation of hierarchical clustering algorithms for document datasets*, en, in: Proceedings of the eleventh international conference on Information and knowledge management, 2002, pp. 515–524.