Journal of Universal Computer Science, vol. 24, no. 5 (2018), 622-633
submitted: 23/1/18, accepted: 28/4/18, appeared: 28/5/18 © *J.UCS*

Advanced Analysis of Data Streams for Critical Infrastructures Protection and Cybersecurity

Barbara Bobowska
(Wroclaw University of Science and Technology, Poland
barbara.bobowska@pwr.edu.pl)

Michał Choraś
(UTP University of Science and Technology in Bydgoszcz, Poland
chorasm@atr.bydgoszcz.pl)

Michał Woźniak
(Wroclaw University of Science and Technology, Poland
michal.wozniak@pwr.edu.pl)

Abstract: Cyber threats are nowadays a major danger to critical infrastructures and to homeland security. For several years now, the focus have been targeted at the physical protection of critical infrastructures. Currently, experts realize that the critical infrastructure can be also attacked via the application layer of computer networks. In order to efficiently protect such critical systems, the huge amount of data has to be efficiently analyzed and correlated. Therefore, this paper focuses on the overview of the advanced data stream processing methods to be applied in the domain of cybersecurity and critical infrastructure protection. The major contribution of this work is the analysis of such innovative aspects as concept drift analysis deployed as the pre-processing step dedicated for anomaly detection systems to counter cyber attacks. Moreover, we discuss the different challenges in data streams analysis including data imbalance and provide solid reasoning why applying a concept drift detector is crucial when designing a modern cybersecurity systems.

Key Words: cybersecurity, machine learning, data science, concept drift, data stream, anomaly detection, data imbalance
Category: I.5, C.2.0

1 Introduction

Security experts and societies have realized long ago that critical infrastructures are vulnerable to many threats and need to be well protected. In recent decades, it became clear that those threats are not only physical and natural, but may also come from the cyber space. Currently, cybersecurity of critical infrastructures is an essential part of national cybersecurity and homeland security strategies, e.g. among EU member states. The general overview and comparison of such strategies is collected in [Alliance, 2016]. The report states that most countries realize that cybersecurity of critical infrastructure should be a national priority. For example National Polish Cybersecurity Strategy for 2017-2022, includes sub-goal, focusing on increased security in context of critical services and ICT infrastructure [Polish Government, 2017]. Similarly, the UK National Cybersecurity Strategy [UK Government, 2016] and French National Digital Security Strategy [Enisa Europe, 2015] address cyber protection of information systems and critical infrastructures. Still, as stated in [Alliance, 2016], the implementation of national strategies vary greatly in different Member States (e.g. legal frameworks, operational details etc.), and that cooperation of nations with private and non-governmental organizations should be improved. Similar aspects and the needed focus on cyber protection of critical infrastructures were postulated by cybersecurity roadmaps, published by EU projects (FP7 CAMINO, FP7 COURAGE, FP7 CyberRoad) [Akhgar et al., 2016].

The importance of cybersecurity in Critical Infrastructure Protection (CIP) is manifested in the current and future research directions in Europe. It is well reflected in the European Commission research efforts, e.g. Horizon 2020 framework programme. Examples of topics addressing cybersecurity in CIP include among others: (a) CIP-01-2016-2017: Prevention, detection, response and mitigation of the combination of physical and cyber threats to the critical infrastructure of Europe (2016-2017 Critical Infrastructure Protection call), (b) SU-INFRA01-2018-2019-2020: Prevention, detection, response and mitigation of combined physical and cyber threats to critical infrastructure in Europe (2018-2020 Critical Infrastructure Protection call), (c) SU-DS04-2018-2020: Cybersecurity in the Electrical Power and Energy System (EPES): an armour against cyber and privacy attacks and data breaches (2018-2020 Digital Security call), (d) SU-DS05-2018-2019: Digital security, privacy, data protection and accountability in critical sectors (2018-2020 Digital Security call).

To date, many works have been devoted to the cyber protection of the lower network levels, since those are closer to the physical processes (e.g. SCADA at power plants). However, there are various threats and injection points, as presented in Fig. 1. In fact, in our opinion, more attention should be targeted at the application layer [Andrysiak et al., 2014, Kozik et al., 2016], and the effective events correlation (e.g. in various layers or from different sensors and probes) [Choraś and Kozik, 2011, Choraś et al., 2011, Choraś et al., 2013]. As presented in Fig. 1 even remote command and control centers are connected to physical processes, and by the successful breach or injection, the critical infrastructure may be threatened, as it already happened in Ukraine (Ivano-Frankivsk region) in December 2015.

The attackers infected the main servers controlling the electricity distribution process, infiltrated in the victims network (possibly using a malware backdoor) and issued a command to open breakers of various substations. Using macro script in Excel files to drop the malware, the infected Excel spreadsheets have been distributed during a spear-phishing campaign, that targeted IT staff and system administrators working for multiple companies, responsible for distributing electricity throughout Ukraine [Kozik et al., 2015, Choraś et al., 2016].

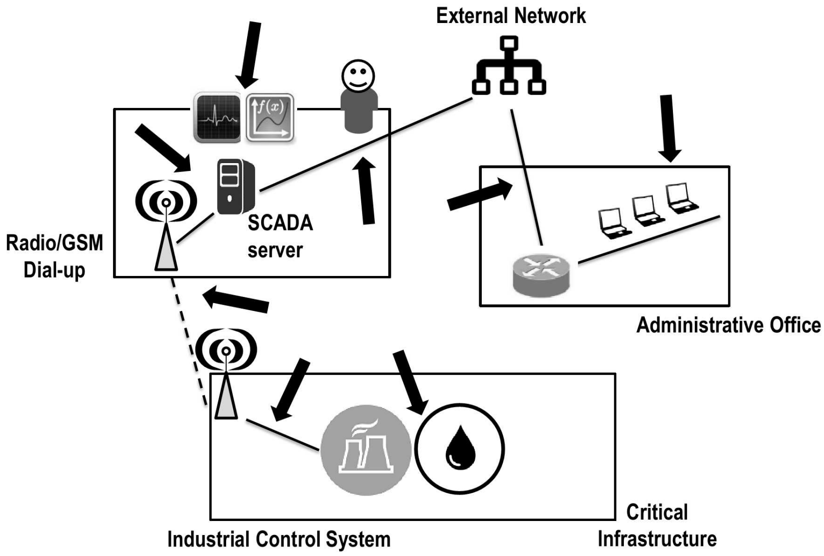

Figure 1: Potential injection points in critical infrastructures

In order to efficiently protect critical infrastructures from cyber attacks, the holistic protection (network levels, remote applications etc.) has to be implemented. This implies the need for advanced data stream processing and correlation. Such situation and challenges in network security motivate our research and the approach to apply concept drift detectors and the lifelong learning approach to cybersecurity domain [Choraś et al., 2017, Choraś and Woźniak, 2017].

Therefore, in this paper, the overview of the advanced data stream processing methods to be applied in the domain of cybersecurity and critical infrastructure protection is presented. The major contribution of this paper is the analysis of such innovative aspect, as concept drift analysis, deployed as the pre-processing step dedicated for anomaly detection systems to counter cyber attacks. We discuss the different challenges in data streams analysis including data imbalance, and provide solid reasoning why applying a concept drift detector is crucial when designing a modern cybersecurity systems.

The paper is structured as follows: in Section 2 the overview of data stream processing is given. In Section 3 the notion of concept drift is introduced, and the proposal of applying concept drift to cybersecurity of critical infrastructures is discussed. Future work and conclusions are given thereafter.

2 Advanced Data Stream Processing

Data stream analysis has gained significant focus from the researchers in last few years due to variety of real-life applications where it can be used. Those include anomaly detection, fraud transaction identification, spam filtering etc. Data streams can be divided into two categories, given the characteristics of processed stream: stationary and non-stationary data streams [Krawczyk et al., 2017]. Analysis of both poses a challenge due to the fact, that it needs to be performed in a timely fashion and that delays should be avoided, given the fact, they could negatively impact the accuracy of the predictions. Secondly, the data arrives in a sequential manner creating potentially infinite amount of data. Given the limited computational and memory resources, the objects cannot be stored in memory and each of the instances are processed only once, before they are discarded making their re-evaluation impossible. For that reason, the information about objects is replaced by statistics. Being able to find the balance between remembering crucial information for model update and storage limitations is of utmost importance. Another issue is the fact that the observer has no influence over how the objects arrive, because there is no prior knowledge about the probability distributions. However, for stationary streams that probabilities (i.e., parameters of their distributions), are fixed meaning the concept is stable. In the latter, either the *prior* probabilities of classes, or class conditional probabilities change over time, resulting in the possibility of the *posterior* probabilities of classes of objects altering. This phenomenon is called *concept drift* and will be further discussed in the upcoming section [Krawczyk et al., 2017] [Krawczyk, 2016]. It is worth noticing that most of the real-life applications are problems from the non-stationary domain. For instance clients purchases which may wary due to the current weather conditions (like certain clothing articles or medicaments that may be more popular during certain season, i.e., winter), and the weather prediction itself, as well as analysis of currently trending topics on social platforms ex. Twitter etc. [Wang and Jones, 2017]

Similarly cyber attacks and network intrusion detection also fall into the category of non-stationary data streams, due to the non-stationary nature of network traffic, and the fact that even slight alteration of the pattern of an attack can make the identification of it impossible for the current signature-based detection systems. Another factor that affects greatly the complexity of the non-stationary data stream analysis is a situation where objects from one particular class are represented by much less instances then objects of other classes. Class imbalance is a commonly present phenomenon, where the class distribution is skewed. This can significantly decrease the quality of a predictive ability of a system, if not approached accordingly, due to the fact that most classifiers are biased towards the majority cases, and as a result ignore the objects from the minority class. Such a situation can cause significant degradation of the system, since more often than not it is the misclassification of the minority class examples that can be much more costly and important, i.e., detecting fraudulent transaction. Other factors contributing to the difficulty of the classification are the overlapping between the classes or a case where a significant number of instances from the minority class is present inside the majority class region as well, as the fact that instances of the minority class often from unstructured clusters (see. Fig. 2) [Napierala and Stefanowski, 2016].

Network anomaly detection is clear example of an imbalanced data problem due to the fact that some transactions are much less frequent than the other while having considerable size [Wang and Jones, 2017].

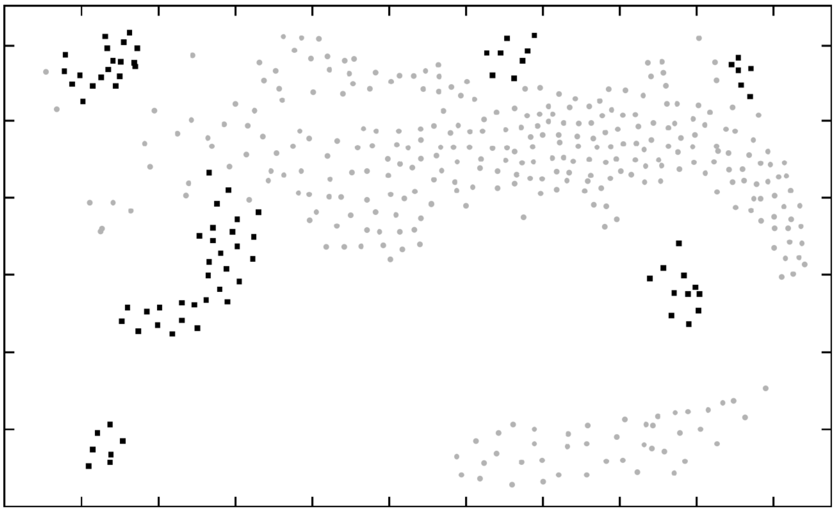

Figure 2: Visualization of an imbalanced dataset

There are three main approaches to imbalanced data classification:

- Data pre-processing methods: the objective of these methods is to handle the data imbalance by evening out the quantity of objects in analyzed classes. This is achieved by either *undersampling* or *oversampling*. In *undersampling* the objects from the majority-class are removed either randomly or using the neighbor analysis. For oversampling the new instances are an artificially generated objects or obtained as a result of random replication of instances from the minority-class. Unfortunately re-sampling the data may have an negative impact on the model i.e. result in *overfitting*, [Krawczyk, 2016] [Hoens et al., 2012].
- Inbuilt mechanisms: concentrate on adapting existing classification algo-

rithms. A cost-sensitive approach, where different costs are assigned to training examples of both minority- and majority-classes, and methods using one-class classification are most common. Alas, such solutions pose difficulty in correctly assigning the costs to the examples as well as creating a reverse bias towards the minority-class [Krawczyk, 2016], [Hoens et al., 2012].

– Hybrid methods: are a combination of aforementioned approaches (e.g. ensembles of classifiers with integrated undersampling and oversampling methods [Krawczyk, 2016], [Hoens et al., 2012]).

3 Concept Drift for cybersecurity of Critical Infrastructures

Concept drift is a change in the probabilistic characteristics of the data stream [Krawczyk et al., 2017]. It means that the characteristics of the decision attributes and of the classes to be predicted, change in time in unpredictable manner. Such situation may cause the decrease in performance of a classier. What is interesting, the accuracy may further decrease with each new portion of the data (in contrast to situation where new data should increase the classification quality). In cybersecurity context, it would mean the decrease of cyber attacks detection in time, due to the natural (not malicious) and unpredictable changes of network traffic characteristics (see. Fig. 3).

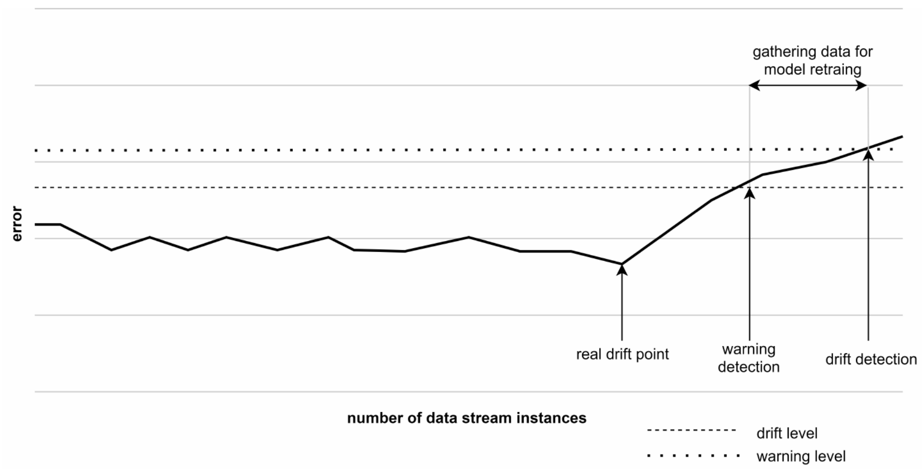

Figure 3: The idea of drift detection based on tracking classifier errors [Krawczyk et al., 2017]

Therefore, in our opinion, the *concept drift* detection techniques, should be applied to the autonomous detection of the model's parameter changes, related to incoming new traffic (and new learning tasks). When an cyber attack is launched its pattern is defined and studied by the network security specialists, in order to design a defense against it. But when the strategy of said attack is changed, that methods are no longer useful. A concept drift can be used to describe such instances. Concept drift [Hoens et al., 2012] may come in many forms, depending on the type of change. Usually, its appearance spoils quality of used models, therefore, developing such methods which can effectively deal with this phenomenon is still a focus of intense research. There are a few taxonomies of concept drift, but lets focus on two of them. One can categorize drifts according how they impact the probabilistic characteristics of a classification task [Gama et al., 2014], i.e.: *virtual concept drift* and *real concept drift.* The former defined as a change in the distribution of feature values does not have any impact on decision boundaries [Widmer and Kubat, 1993]. *real concept drift* resulting from the changes in class conditional probabilities, is a modification in posterior probabilities of the classes, and in consequence has an impact on decision boundaries. While detection of the *real concept drift* is most important, recognizing the virtual one may be also useful considering a case where a drift detection causes model rebuilding, an action not necessary when detecting virtual drift. It is important to note, that while the objective is to rebuild the model whilst ensuring its quality remains satisfactory, that alternation causes delay, in the form of restoration time which should be as short as possible (see. Fig. 4).

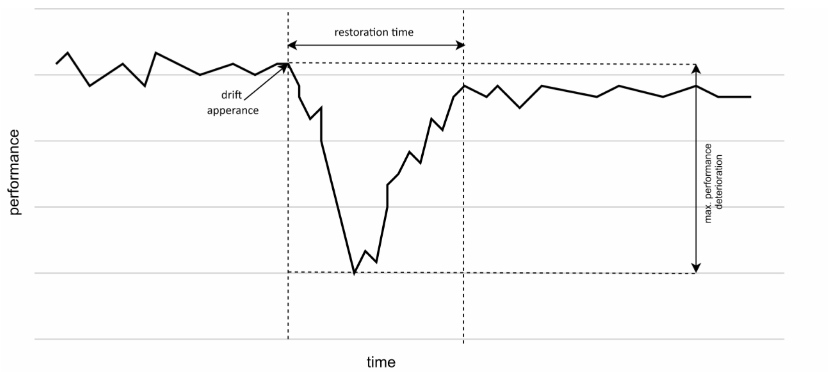

Figure 4: The illustration of the model restoration time [Krawczyk et al., 2017]

We may also distinguish the drift types according to the drift impetuosity, i.e., (i) for the *gradual drift* (see. Fig.5a) for a given period of time, examples from different models may appear in the stream concurrently, while for the *incremental drift* (see. Fig.5b) the model's parameters are changing smoothly and (ii) *sudden* *drift*, where the drift has rapid nature (see Fig.5c) and sometimes the outdated model may appear again (see. Fig.5d). In the domain of cybersecurity a sudden concept drift is equivalent to a new type of attack or a considerable change in its strategy, while a gradual concept drift describes a situation where some cyber criminals use an old version of an attack while a new one has emerged.

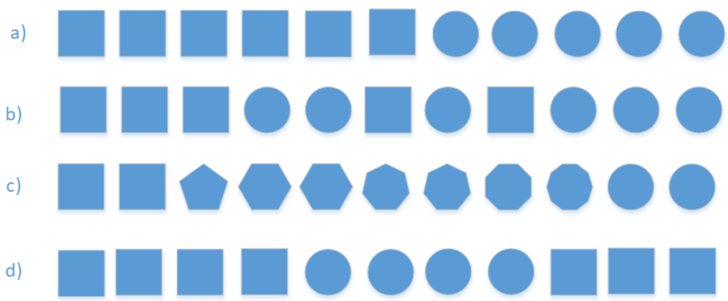

Figure 5: Types of changes in the data stream: (a) sudden, (b) gradual, (c) incremental, (d) recurring

There are several ways to handle concept drifts, that can be generally divided into two categories: i) methods where the learner is adapted at regular intervals, whether the changes have actually occurred or not ex. time windows of fixed size, where choosing an appropriate size of the window is the main challenge, or weighted examples, where an idea that the importance of examples decreases with time (in some cases only the most recent examples are used to rebuild the model) is applied; ii) methods where the learner is updated only if the changes in concept are actually detected ex. if a concept drift is detected, an action to adapt to those alternations is performed. For instance adjusting the window size accordingly to the expanse of concept drift.

Most of concept drift detectors use information about the performance convergence of the operated predictive model to return signal, if probability distributions are changing. Usually, they can return the signal that drift is detected and model should be rebuilt as quick as possible, or that the so-called warning level is achieved, which may cause the necessity to collect new data to rebuild/update the model. The drift detector may be recognized as a classifier, but it rather solves a regression problem, i.e., evaluating how far the characteristics (as probability distributions) of operated model are from the characteristics describing the real problem under consideration. Such a task is tough, because on one hand we should detect a drift as soon as possible to replace outdated model and to reduce so-called restoration time, but on the other we do not accept too many false alarms [Gustafsson, 2001]. Additionally, it is worth noticing that drift detectors are usually assuming the continue access to class labels, which cannot be granted from the practical point of view. Therefore, building such systems we should take into consideration the data labeling cost, which is usually passed over. Unfortunately, without access to class labels the real drift could be undetected [Sobolewski and Woźniak, 2013].

Cyber attacks detection methods and systems can be as weak, as weak are the advanced data processing approaches and algorithms. The standard approach to cyber attacks detection is the so-called signature-based mode, where the patterns of evil malicious traffic are compared to current traffic samples, and if matched, the alarm is raised (Fig. 6). Such approach is, of course, inefficient in detection of the new or modified cyber attacks, or so-called 0-day exploits. Therefore, another approach is to detect anomalies. To do so, firstly, the pattern of normality (normal traffic etc.) has to be learnt and then matched versus the current traffic samples. Whenever, there is no match, the alarm is raised (Fig. 7). This approach is however plagued by false positives (false alarms). Quite often, when the characteristics of network traffic (or e.g. HTTP requests in the application layer) change, such situation is detected as anomaly, even though it is just a normal change of the network behavior. In pattern recognition, such situations are termed as concept drift, and this aspect should be taken into account in detection systems. Therefore in this paper, we postulate to include concept drift detector, as a form of a pre-processing step in anomaly detection cybersecurity systems (Fig. 8).

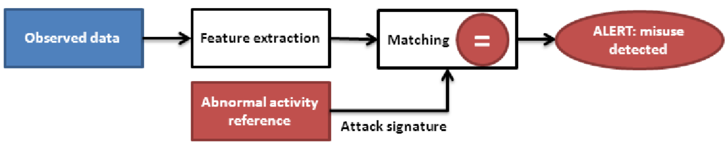

Figure 6: Signature-based approach for cyber attacks detection

4 Future Work and Conclusions

We currently focus on the adaptation of the concept drift detection mechanism in the area of cybersecurity. In particular, we plan to adapt the above-mentioned ideas to efficiently detect anomalies in the monitored networks. Application layer attacks, such as SQLIA (SQL Injection Attacks), are top-ranked on several threat lists. One example of such is the "OWASP Top 10 [OWASP, 2018] list, that

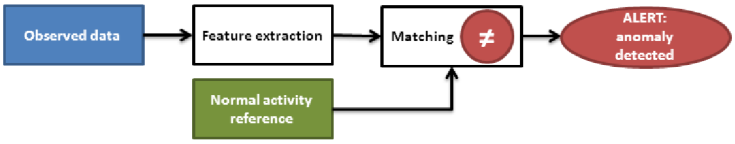

Figure 7: Anomaly detection approach for cybersecurity detection

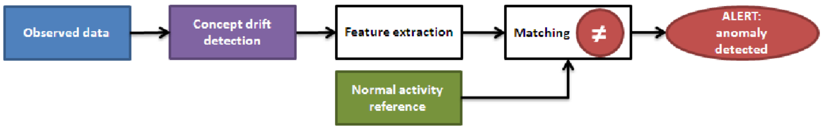

Figure 8: Anomaly detection approach enhanced with the concept drift detection

has been identified by Open Web Application Security. The list, among others, contains the following items from the application layer: injection flaws (e.g. SQL Injection), broken authentication and session management, Cross-Site Scripting. The practical implementation of concept drift and anomaly detection (as in Fig. 8) approach, to protect the application layer, can consider for example the analysis of user requests to web service or data-base. In such scenario, we can apply sensors and implement complex algorithms, to learn the models of normal requests or user behavior, and detect all the requests that fall outside the model of normality. However, in order to decrease the false positive ratio (indicating anomalies which are not attacks or symptoms of misbehavior), we now work on implementing the concept drift detectors as well as applying the lifelong learning solutions. In such scenario, the model of normal requests changes quickly (e.g. due to availability of new services, or new fields in web forms), and therefore anomaly detection approach tends to have high false positive ratio. In our proposed solution, the system will:

– Detect concept drifts and react to them,

– re-learn using the past knowledge to quickly adapt to network changes,

in order to effectively detect and counter cyber atacks on critical infrastructures.

The security of modern computer systems and networks is among the most pressing and discussed matters, that the modern society has to face. Given that currently used approaches, based predominantly on signature based methods have proven to be ineffective against the emerging attacks and malware, designing new methods, that would be able to identify the continuously evolving, intricate cyber attacks, is of utmost importance. Data stream analysis has gained much popularity due to the many real-life applications where it can be utilized. One of such is network anomaly detection among many others. Data stream analysis is a challenging area of research, owing to the amount of data that needs to be analyzed and its often changing characteristics, on top of the computational and memory resources constraints. The task is complicated greatly as a consequence of the existence of such phenomena like concept drift and class imbalance. Another difficulties, such as class overlapping, or the irregularly shaped clusters of examples from minority class, must also be addressed. Concept drifts can cause a significant degradation in performance of the classifier, and therefore in detection of the cyber attacks on time. It's important to note that when implementing a concept drift detector, one has to consider the trade-off between the detection delay and the quality of its performance. Consequently a correct detection of concept drift is a milestone in creating a cybersecurity lifelong learning intelligent system.

Acknowledgement

This work was supported by the Polish National Science Center under the grant no. UMO-2015/19/B/ST6/01597.

References

[Akhgar et al., 2016] Akhgar, B., Choraś, M., Brewster, B., Bosco, F., Vermeersch, E., Luda, V., Puchalski, D., and Wells, D. (2016). *Consolidated Taxonomy and Research Roadmap for Cybercrime and Cyberterrorism*, pages 295–321. Springer International Publishing, Cham.

[Alliance, 2016] Alliance, B. T. S. (2016). *EU Cybersecurity Dashboard. A Path to a Secure European Cyberspace.*

[Andrysiak et al., 2014] Andrysiak, T., Saganowski, Ł., Choraś, M., and Kozik, R. (2014). Network traffic prediction and anomaly detection based on arfima model. In de la Puerta, J. G., Ferreira, I. G., Bringas, P. G., Klett, F., Abraham, A., de Carvalho, A. C., Herrero, Á., Baruque, B., Quintián, H., and Corchado, E., editors, *International Joint Conference SOCO'14-CISIS'14-ICEUTE'14*, pages 545–554, Cham. Springer International Publishing.

[Choraś and Kozik, 2011] Choraś, M. and Kozik, R. (2011). Network event correlation and semantic reasoning for federated networks protection system. In Chaki, N. and Cortesi, A., editors, *Computer Information Systems – Analysis and Technologies*, pages 48–54, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Choraś et al., 2016] Choraś, M., Kozik, R., Flizikowski, A., Hołubowicz, W., and Renk, R. (2016). *Cyber Threats Impacting Critical Infrastructures*, pages 139–161. Springer International Publishing, Cham.

[Choraś et al., 2011] Choraś, M., Kozik, R., Piotrowski, R., Brzostek, J., and Hołubowicz, W. (2011). Network events correlation for federated networks protection system. In Abramowicz, W., Llorente, I. M., Surridge, M., Zisman, A., and

Vayssière, J., editors, *Towards a Service-Based Internet*, pages 100–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Choraś et al., 2013] Choraś, M., Kozik, R., Puchalski, D., and Hołubowicz, W. (2013). Correlation approach for sql injection attacks detection. In *Herrero A. et al (Eds.), Advances in Intelligent and Soft Computing*, pages 177–185, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Choraś et al., 2017] Choraś, M., Kozik, R., Renk, R., and Hołubowicz, W. (2017). *The Concept of Applying Lifelong Learning Paradigm to Cybersecurity*, pages 663–671. Springer International Publishing, Cham.

[Choraś and Woźniak, 2017] Choraś, M. and Woźniak, M. (2017). *Concept Drift Analysis for Improving Anomaly Detection Systems in Cybersecurity*, pages 35–42. University of Maribor Press, Cham.

[Enisa Europe, 2015] Enisa Europe (2015). French national digital security strategy. `https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map/France_Cyber_Security_Strategy.pdf/at_download/file`.

[Gama et al., 2014] Gama, J. a., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37.

[Gustafsson, 2001] Gustafsson, F. (2001). *Front Matter and Index*, pages i–x. John Wiley & Sons, Ltd.

[Hoens et al., 2012] Hoens, T. R., Polikar, R., and Chawla, N. V. (2012). Learning from streaming data with concept drift and imbalance: an overview. *Progress in Artificial Intelligence*, 1(1):89–101.

[Kozik et al., 2015] Kozik, R., Choraś, M., Flizikowski, A., Theocharidou, M., Rosato, V., and Rome, E. (2015). Advanced services for critical infrastructures protection. *Journal of Ambient Intelligence and Humanized Computing*, 6(6):783–795.

[Kozik et al., 2016] Kozik, R., Choraś, M., Renk, R., and Hołubowicz, W. (2016). *Cyber Security of the Application Layer of Mission Critical Industrial Systems*, pages 342–351. Springer International Publishing, Cham.

[Krawczyk, 2016] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

[Krawczyk et al., 2017] Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., and Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132 – 156.

[Napierala and Stefanowski, 2016] Napierala, K. and Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *J. Intell. Inf. Syst.*, 46(3):563–597.

[OWASP, 2018] OWASP (2018). Owasp the open web application project owasp top ten.

[Polish Government, 2017] Polish Government (2017). Strategia cyberbezpieczenstwa rzeczypospolitej polskiej. `https://www.gov.pl/cyfryzacja/strategia-cyberbezpieczenstwa-rzeczypospolitej-polskiej-na-lata-2017-2022`.

[Sobolewski and Woźniak, 2013] Sobolewski, P. and Woźniak, M. (2013). Concept drift detection and model selection with simulated recurrence and ensembles of statistical detectors. 19(4):462–483. `http://www.jucs.org/jucs_19_4/concept_drift_detection_and`.

[UK Government, 2016] UK Government (2016). National cyber security strategy 2016 to 2021. `https://www.gov.uk/government/publications/national-cyber-security-strategy-2016-to-2021`.

[Wang and Jones, 2017] Wang, L. and Jones, R. (2017). Big data analytics for network intrusion detection: A survey. *International Journal of Networks and Communications*, 7(1):24–31.

[Widmer and Kubat, 1993] Widmer, G. and Kubat, M. (1993). *Effective learning in dynamic environments by explicit context tracking*, pages 227–243. Springer Berlin Heidelberg, Berlin, Heidelberg.