A survey on the use of data points in IDS research

Heini Ahde, Sampsa Rauti, and Ville Leppanen

University of Turku, Finland
sampsa.rauti@utu.fi

Abstract. In today's diverse cyber threat landscape, anomaly-based intrusion detection systems that learn the normal behavior of a system and have the ability to detect previously unknown attacks are needed. However, the data gathered by the intrusion detection system is useless if we do not form reasonable data points for machine learning methods to work, based on the collected data sets. In this paper, we present a survey on data points used in previous research in the context of anomaly-based IDS research. We also introduce a novel categorization of the features used to form these data points.

Keywords: Network security, intrusion detection, data points

1 Introduction

An *intrusion detection system* (IDS) is a system used to detect malicious activities in a specific network or system. A network intrusion detection system (NIDS) is placed in the network to monitor network traffic and to detect malicious activities by recognizing anomalous patterns in the incoming or outgoing packets. [1–3, 11] NIDS systems have existed already for quite a long time, and now host intrusion detection systems (HIDS) are becoming increasingly popular. Compared to NIDS, one can gather much more data in HIDS systems, because the system can detect all the internal events related to the inner workings of an operating system and its processes.

Signature-based IDSs detect attacks by observing previously recorded malicious patterns, such as malicious sequences of instructions. Anomaly-based IDSs, on the other hand, use machine learning to learn what normal activity in the network looks like and compare detected behavior to this profile. The strength of this approach is the ability to detect previously unknown attacks. [1, 11, 21] Today, anomaly-based intrusion detection is needed to deal with the increased amount of new variations of malicious programs and targeted attacks [7].

However, the data gathered by the system is useless if we cannot use it correctly to learn about the normal behavior [21]. We have to make a choice about which features in the data we are going to look at and how they are used in our analysis. The same is also true when we are configuring an IDS and building rules for it; we have to know what kind of features we are looking for.

Specifically, it is very important to carefully choose the data points we use to draw conclusions from the data.

A *data point* is a single observed unit of information, a measured collection of features. For example, a simple data point in NIDS research could be a size of a specific network packet. On the other hand, a data point can be a more complex combination of several features, such as combining the IP addresses and port numbers of the sender and the receiver.

The contributions of this paper are as follows. In this survey, we review different types of data points used in literature in the context of anomaly-based IDSs. We look at how these data points have been aggregated from the gathered data and for what kind of purposes they have been utilized. We also provide a novel categorization of different features used to form the data points in IDS research. The categorization presented in this study can be used when choosing a meaningful set of features for data points in NIDS systems and research.

This research is a part of a project where we have had a change to utilize a new commercial HIDS system by F-Secure to collect a lot of data on internal events with the aim of creating new kind of machine learning solutions for (ab)normal behaviour profiling. Moreover, we are working towards an open-source implementation of system enabling one to collect data related to internal events and thus build such systems. As such systems produce a vast amount of data, efficient machine learning based methods are needed. Therefore, the purpose of this paper is to survey what kinds of features have been used in this context in the existing literature.

The rest of the paper is organized as follows. Section 2 presents the study setting and the research questions we aim to answer in this study. Section 3 presents a survey on the types of features used to form data points in NIDS research. Section 4 presents our classification for different types of features in NIDS research and their usual purposes. Section 5 concludes the paper.

2 The study setting and the research questions

The data collected by an IDS often consists of a large amount of events, each having lots of features or attributes. One approach would be to feed all of this data to a machine learning algorithm as such. However, the setting would easily become very complex as the data points would have a high number of dimensions.

Although one can use a multidimensional dataset to create profiles, this approach is often ineffective. The dataset often contains several irrelevant features that do not have a significant effect on the end result. Moreover, handling high dimensional datasets obviously increases the calculation time. Hotho et al. [10] note that a low-dimensional space is more favorable for finding clear clusters in the data.

It is often worth considering more elementary approaches and only looking at narrow slices of data. As a practical example, such a narrow a view could be a set of day-wise or machine-wise points formed from port numbers used by a selected application.

Because of the obvious need to choose good and meaningful data points for low-dimensional datasets, we have formulated the following two research questions for this study:

RQ1 What are the types of data points used in NIDS research?

The first research question studies data points in NIDS research considering statistical and machine learning based methods, and how these data points have been aggregated from the data (which features are included?).

RQ2 For what kind of purposes are the data points used?

The second question inquires what kind of observations the discovered data points aim to derive from the data.

In what follows, we present a review of the types of data points proposed and used in network intrusion detection systems. The purpose is to find out what kind of different network-related data points exist and for what purposes they have been utilized. To search for relevant publications, we used Google Scholar and keywords "profiling network traffic", "data mining network traffic" and "internet traffic behaviour profiling for network security monitoring" to collect literature references. Moreover, snowballing (looking at the references of found papers) was performed to make the search more complete. The publications that did not discuss aggregation of data points in NIDS research were excluded.

3 Types of network traffic data points

Shadi et al. [18] presented a hierarchical clustering algorithm for network flow data. They generated their data points in a 2D space of source and destination IP addresses. This 2D space represented all possible values for the source and destination IP addresses. In a related studies, Estan et al. [5] introduced a multidimensional traffic clustering, for analyzing IP-based traffic. Their method can cluster traffic along multiple different dimensions including source and destination IP addresses, protocol, source and destination port numbers. Mahmood et al. [15] introduced a new clustering method for generating conclusions of significant traffic flows. The key attributes were source and destination IP addresses, protocol, source and destination port numbers.

Xu et al. [25] studied significant behavior patterns for network security monitoring (detecting anomalies). They collected Internet backbone traffic flow and constructed four collections of clusters based on following extracted features:

- The source IP address
- The destination IP address
- The source port number
- The destination port number

Feroz et al. [6] demonstrated that cluster labels increase the classifier accuracy. They presented an approach that classifies URLs based on their lexical and host-based features. In a related study, Ma et al. [13] extracted lexical and host-based features (such as IP address of the URL, the mail exchanger and the name server) from URLs.

Packet header traces are widely used in data mining and machine learning analysis. McGregor et al. [16] used the following features and expectation-maximisation clustering algorithm to group the traffic flows into a small number of clusters:

- Byte counts
- Connection duration
- Interarrival statistics
- Packet size statistics (minimum, maximum, quartiles, minimum as fraction of max and the first five modes)
- The number of transitions between transaction mode and bulk transfer mode (the time when there was more than three successive packets in the same direction without any packets carrying data in the other direction)
- The time spent and the idle (all periods of 2 seconds or greater when no packet was seen in either direction)

Liu et al. [12] examined supervised and unsupervised machine-learning techniques to classify network traffic by TCP-based applications. Their K-means approach took following statistical information as an input vector to build classifiers (a = client, b = server):

- The number of total-packects-b-a
- The number of actual-data-bytes-b-a
- The number of pushed-data-pkts-b-a
- Size of the mean-IPpacket-a-b and size of the mean-IPpacket-b-a
- Size of the max-IPpacket-a-b and size of the max-IPpackect-b-a
- Variant of the IPpacketsize-a-b and variant of the IPpacketsize-b-a
- Duration

Magdalinos et al. [14] examined on how to automatically create user profiles and how to use these profiles to enhance the performance of network control functions. They applied well-known data mining and machine learning tools, such as k-means and naive Bayes, and following *context extraction and profiling engine* (CEPE) information:

- The status of the user (name, age, gender, education, operating systems, screen width and height, etc.)
- The device (time, transmit power, lost packets etc.)
- The combination of the service type (web, ftp, video, etc.) and network information (lost packets, transmit power, cell type, etc.)
- A log file (all monitored parameters)

Erman et al. [4] applied both supervised and unsupervised machine learning methods to classify network traffic. The authors generated data points using statistical features, such as total number of packets, mean data packet size, flow duration and the mean inter-arrival time of packets.

Hammerschmidt et al. [9] built communication profiles using connection-level IP flow records, such as transport protocol of the flow, time since previous flow started, duration of the flow, count of packet exchanged and amount of data received. One of their main purpose was to extract the key behavior from the records and also reduce irrelevant behavior. They learned communication profiles with the *DFASAT software package* using an IP flow dataset that contains real communication from hosts running botnet malware as well as legitimate traffic.

Wang et al. [23] applied two benchmark datasets, 20Newsgroups (20NG) and RCV1, to evaluate domain dependent document clustering. 20NG contains 20000 newsgroups documents and RCV1 contains manually labeled newswire stories from Reuters Ltd.

Gonzalez et al. [8] presented a platform *Net2Vec*, that is able to capture raw network flow data, transform it into meaningful data points and apply the predictions over the data points in real time. To showcase the applicability of the Net2Vec they constructed a user profiling scheme. They generated a two-element data point consisting of source IP address and hostname.

Singh et al. [20] examined anomaly based intrusion detection systems that learn to distinguish normal behaviour and abnormal behaviour. They used *NLS-KDD dataset* and *Kyoto University dataset*. The features of NLS-KDD dataset included for example duration, protocol type, byte information, service information, and log information.

Sarmadi et al. [17] examined the feasibility of profiling internet users based on volume and time of usage. As an experimental dataset they used real internet usage data collected via NetFlow logs (metadata). The data was collected in one month from 66 university students. The traffic flow contains information about usage time, octets, packets, port numbers and protocols. However, the focus of the paper was to examine only the following combination:

- Duration (the amount of milliseconds from the start of the flow to the end)
- Octets (the number of layer 3 bytes of the flow)

4 A categorization for data point features

To find a meaningful categorization for the data point features reviewed in the previous section, internet traffic classification methods presented in the previous literature can give us some clues. Wang et al. [24] and Singh [19] divide the internet traffic classification methods to following groups:

- *Port-based internet traffic classification.* In this classification method, port numbers in headers of the transport layer protocol, such as TCP, are examined and used when forming data points.

Table 1. Distribution of NIDS data point related research within our feature categorization.

Category	**Example features**	**Papers**
Network addresses and ports	IP addresses, ports, hostnames, domains	[18, 6, 8, 25, 5, 15, 13]
Protocols and service types	HTTP, voice, voip, video	[5, 15, 20, 9]
Statistical features	connection duration, packet size, idle time	[16, 17, 12, 4, 14, 20, 9]
Network information	lost packets, delay log information	[14, 20]
External features	screen type, user's education	[23, 14]

- *Payload-based internet traffic classification.* This approach, also called deep packet inspection, inspects packet payloads for characteristic signatures of known applications. It can be used to check that the content is supplied in the correct format and to make sure payload is not malicious.
- *Classification based on statistical traffic properties.* Statistical characteristics of internet traffic at the network layer, such as packet length, flow duration, inter arrival time of packet, standard deviation, are used to classify traffic.
- *Internet traffic classification using machine learning.* A dataset consisting of number several data points (that in turn consist of one or several features or attributes) is created. The output will be some specific pattern discovered in the data.

A similar taxonomy is used by Valenti et al. [22], as they also list port-based classification, payload-based classification, statistical classification as traffic classification techniques. In addition, the authors also include stochastic packet inspection and behavioural traffic classification in their taxonomy. Stochastic packet inspection uses the statistical fingerprints in the application layer headers and uses them to recognize formats of different application protocols. Behavioral classification generally monitors traffic of one host as a whole and examines traffic patterns (such as which transport level protocols are used and how many hosts are contacted), trying to profile applications that are executed on the target host.

The two categorizations do not directly fit for classifying data point features into groups. For example, they seem to exclude many potential types of features that may be important in profiling hosts or specific users, such as IP addresses or information about the user. Also, some categories, such as "machine learning" or "behavioral classification" are too coarse and make no sense in the context of categorizing features (after all, all features can be used for machine learning and behavioral analysis). However, some categories like data related to ports, packet payload or traffic statistics, seem to make more sense for feature types. Our categorization for NIDS research data points is presented in Table 1.

The categorization consists of the following categories:

Network addresses and ports. Network addresses and port defining the hosts and applications at the endpoints are important features when profiling hosts based on network traffic. Port numbers have traditionally been used to identify applications. Recently, however, an increasing number of programs use nonstandard ports, and malicious programs might use well-known protocol ports (such as 80 reserved for HTTP) in order to disguise their presence. Therefore, port numbers are often combined with other features, such as IP addresses [25] or used protocol [5] to form more informative data points. Domain names associated with IP addresses may also be useful as features, for instance when trying to identify malicious websites [13].

Used protocol or service type. The used protocol (such as HTTP, FTP, SMTP...) can be used as a feature to profiling applications and clustering multidimensional network traffic [15]. A feature similar to the used protocol is service type, for example web, video or voip. Information about the protocol and service along with statistical features is useful for creating a model to profile hosts based on their traffic statistics [9] or giving predictions on the expected behavior of end users.

Statistical features. Statistical data point features are related to traffic flow. They include for example statistics of interarrival times, connection durations, packet sizes and number of packets from host a to host b. Statistical features such as mean interarrival time and accumulated idle time are aggregated from other values. Several of these statistical features are often bundled together to form data points, for example in order to build a profiles for applications on a specific host [12].

Network information. Network related features such as the transmit power of the network, network cell related information or delays in the network can also be used in profiling. However, when trying to profile hosts, applications or users, these features are usually combined with features from other categories to form data points.

External features. An external feature refers to a feature not directly related to network traffic or properties of packets. Instead, this category includes data on the users and devices they use. Examples of such features include type of processor, operating system, coordinates of a mobile device, and completely non-technical features such as age and gender of the user. These features are usually used to build profiles for users, like Magladinos et al. [14] do by combining for instance features of users, devices and service types to form data points.

5 Conclusion

In this paper, we have reviewed different types of data points utilized in NIDS research. We have studied what sort of features have been previously used to

aggregate data points in the existing literature. Based on these findings, we proposed a novel categorization for features used to form data points.

A data point is often multidimensional and consists of many features, often from several categories outlined previously. Since clustering high-dimensional datasets is difficult and computationally ineffective, we believe the new categorization we have provided in this study can help in choosing meaningful features for low-dimensional data points. While the features and data points vary in each individual system, we have seen that many feature categories that are used regularly.

In the future research, it might be interesting to experiment how well our classification works when building a NIDS system and classifying data gathered in the system. This would paint a clearer picture of how our classification can be used in NIDS research and how effective it is. The future work also includes reviewing features and data points used in the research related to HIDS systems, as this area grows more mature and including them in our categorization.

References

1. Al-Jarrah, O., Arafat, A.: Network Intrusion Detection System using attack behavior classification. In: 5th International Conference on Information and Communication Systems (ICICS), 2014, IEEE (2014) 1–6
2. Alanazi, H., Noor, R., Zaidan, B., Zaidan, A.: Intrusion detection system: overview. arXiv preprint arXiv:1002.4047 (2010)
3. Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: Network anomaly detection: methods, systems and tools. IEEE communications surveys & tutorials **16**(1) (2014) 303–336
4. Erman, J., Mahanti, A., Arlitt, M.: Qrp05-4: Internet traffic identification using machine learning. In: Global Telecommunications Conference, 2006. GLOBECOM'06, IEEE (2006) 1–6
5. Estan, C., Savage, S., Varghese, G.: Automatically inferring patterns of resource consumption in network traffic. In: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, ACM (2003) 137–148
6. Feroz, M.N., Mengel, S.: Phishing URL detection using URL Ranking. In: Big Data (BigData Congress), 2015 IEEE International Congress on, IEEE (2015) 635–638
7. Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & security **28**(1-2) (2009) 18–28
8. Gonzalez, R., Manco, F., Garcia-Duran, A., Mendes, J., Huici, F., Niccolini, S., Niepert, M.: Net2Vec: Deep learning for the network. In: Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks, ACM (2017) 13–18
9. Hammerschmidt, C., Marchal, S., State, R., Pellegrino, G., Verwer, S.: Efficient learning of communication profiles from IP flow records. In: 41st Conference on Local Computer Networks (LCN), 2016, IEEE (2016) 559–562
10. Hotho, A., Maedche, A., Staab, S.: Ontology-based text document clustering. Künsliche Intelligenz (KI) **16**(4) (2002) 48–54

11. Kemmerer, R.A., Vigna, G.: Intrusion detection: a brief history and overview. Computer **35**(4) (2002) supl27–supl30
12. Liu, Y., Li, W., Li, Y.C.: Network traffic classification using k-means clustering. In: Computer and Computational Sciences, 2007. IMSCCS 2007. Second International Multi-Symposiums on, IEEE (2007) 360–365
13. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2009) 1245–1254
14. Magdalinos, P., Barmpounakis, S., Spapis, P., Kaloxylos, A., Kyprianidis, G., Kousaridas, A., Alonistioti, N., Zhou, C.: A context extraction and profiling engine for 5G network resource mapping. Computer Communications **109** (2017) 184–201
15. Mahmood, A.N., Leckie, C., Udaya, P.: An efficient clustering scheme to exploit hierarchical data in network traffic analysis. IEEE Transactions on Knowledge and Data Engineering **20**(6) (2008) 752–767
16. McGregor, A., Hall, M., Lorier, P., Brunskill, J.: Flow clustering using machine learning techniques. In: International Workshop on Passive and Active Network Measurement, Springer (2004) 205–214
17. Sarmadi, S., Li, M., Chellappan, S.: On the feasibility of profiling internet users based on volume and time of usage. In: 9th Latin-American Conference on Communications (LATINCOM), 2017, IEEE (2017) 1–6
18. Shadi, K., Natarajan, P., Dovrolis, C.: Hierarchical IP flow clustering. In: Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks, ACM (2017) 25–30
19. Singh, H.: Performance analysis of unsupervised machine learning techniques for network traffic classification. In: Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on, IEEE (2015) 401–404
20. Singh, R., Kumar, H., Singla, R.: An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Systems with Applications **42**(22) (2015) 8609–8624
21. Tsai, C.F., Hsu, Y.F., Lin, C.Y., Lin, W.Y.: Intrusion detection by machine learning: A review. Expert Systems with Applications **36**(10) (2009) 11994–12000
22. Valenti, S., Rossi, D., Dainotti, A., Pescapè, A., Finamore, A., Mellia, M.: Reviewing traffic classification. In: Data Traffic Monitoring and Analysis. Springer (2013) 123–147
23. Wang, C., Song, Y., El-Kishky, A., Roth, D., Zhang, M., Han, J.: Incorporating world knowledge to document clustering via heterogeneous information networks. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2015) 1215–1224
24. Wang, Y., Xiang, Y., Zhang, J., Zhou, W., Wei, G., Yang, L.T.: Internet traffic classification using constrained clustering. IEEE transactions on parallel and distributed systems **25**(11) (2014) 2932–2943
25. Xu, K., Zhang, Z.L., Bhattacharyya, S.: Internet traffic behavior profiling for network security monitoring. IEEE/ACM Transactions On Networking **16**(6) (2008) 1241–1252