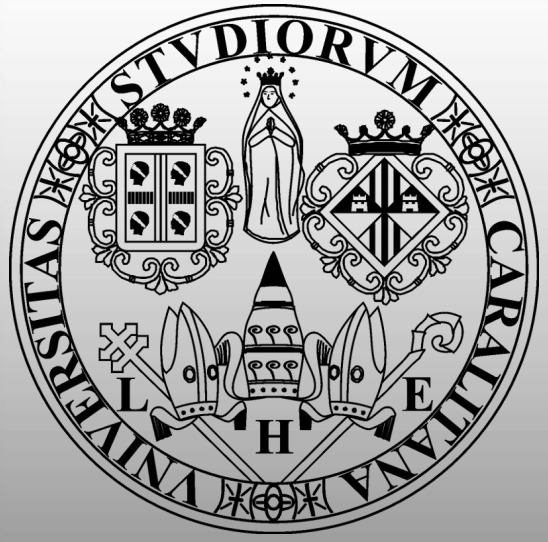


A Discretized Extended Feature Space (DEFS) Model to Improve the Anomaly Detection Performance in Network Intrusion Detection Systems

Roberto Saia, Salvatore Carta, Diego Reforgiato Recupero, Gianni Fenu, and Maria Madalina Stanciu

Department of Mathematics and Computer Science
University of Cagliari, Italy

KDIR 2019, 11th International Conference on Knowledge Discovery and Information Retrieval
17÷19 September 2019 - Vienna, Austria

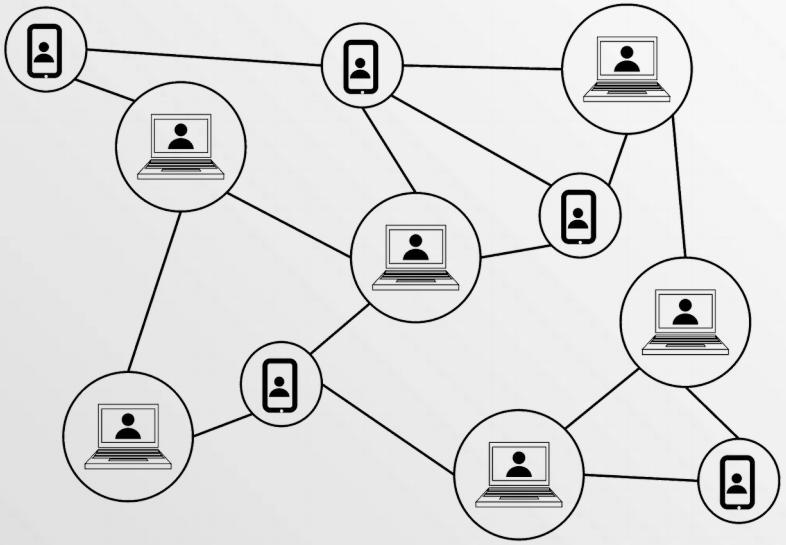
Outline

- **Overview**
  Introduction ▪ State of the Art ▪ Open Problems
- **Intuition**
  Idea ▪ Formalization
- **Implementation**
  Data Discretization ▪ Data Extension ▪ Event Classification
- **Results**
  Experiments ▪ Conclusions and Future Work

OVERVIEW

Introduction

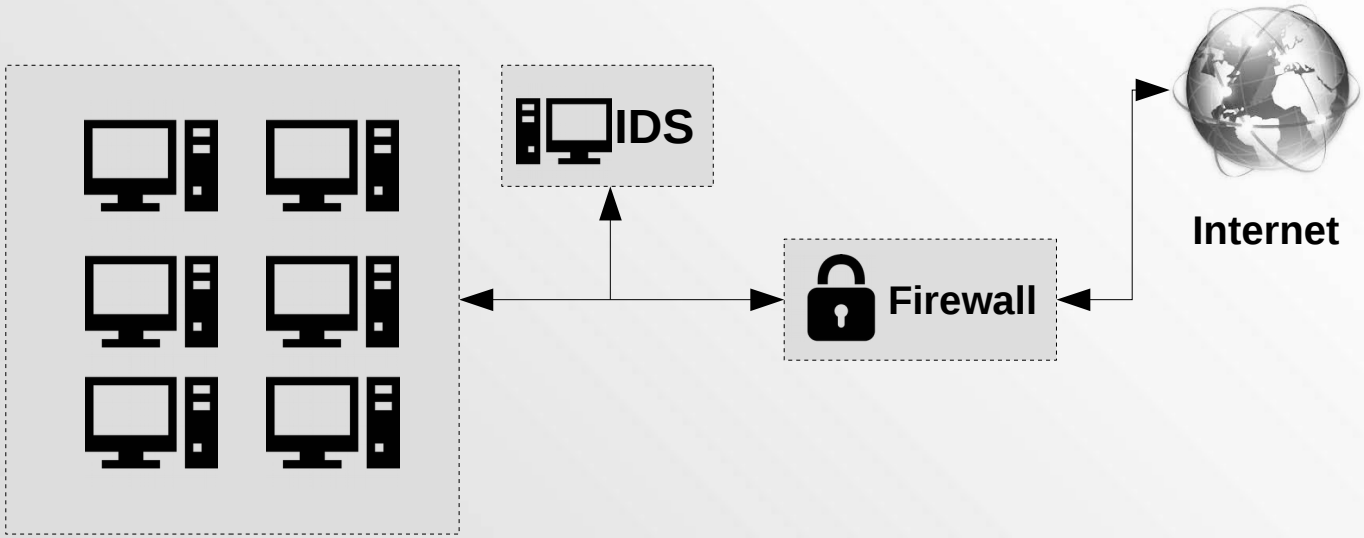
Network security represents one of the most crucial aspects of today's society, in which an ever increasing number of services is performed through networks such as the Internet



Roberto Saia et al. - *A Discretized Extended Feature Space Model to Improve the Anomaly Detection Performance in Network Intrusion Detection Systems*

Introduction

The **Intrusion Detection Systems** (IDSs) are the main instruments to face such a problem, since they have been designed in order to classify the network events into two classes, **normal** or **intrusion**



State of the Art

- Many **approaches** have been used in order to detect anomalous network activities, for instance:
  - *Fuzzy Logic*
  - *Clustering Algorithms*
  - *Support Vector Machines*
  - *Genetic Algorithms*
  - *Hybrid Solutions*

Roberto Saia et al. - *A Discretized Extended Feature Space Model to Improve the Anomaly Detection Performance in Network Intrusion Detection Systems*

Open Problems

Regardless of the adopted technique, the definition of effective evaluation models represents a **hard challenge** for several factors, such as:

- High level of **heterogeneity** and **dynamism** is the involved data
- The **behavior** of an attack can be almost identical to that of a legitimate activity

INTUITION

Idea

- The idea behind the proposed approach relies into the observation that the main limitation of the state-of-the-art solutions is given by their inability to classify the new network events correctly
- In order to face this problem we introduce a new feature space where the event features have been transformed through a discretization and extension criterion

Formalization

- The discretization process is able to reduce the high number of event patterns by grouping the similar ones
- The extension process extends the number of event features by introducing a series of meta-information with the aim to obtain a better event characterization
- The extension process reduces the issues related to the loss of information caused by a canonical discretization process

IMPLEMENTATION

Implementation

Step 1 of 3 - Data Discretization

On the basis of a η value experimentally defined, we perform the **data discretization** process $\xrightarrow{\eta}$

$$\{f_1, f_2, \ldots, f_N\} \xrightarrow{\eta} \{d_1, d_2, \ldots, d_N\}, \ \forall \ e \in E$$

Where f_1, f_2, ..., f_N denotes the original feature values of an event **e**, and d_1, d_2, ..., d_N denotes these values after the discretization process

Implementation

Step 2 of 3 - Data Extension

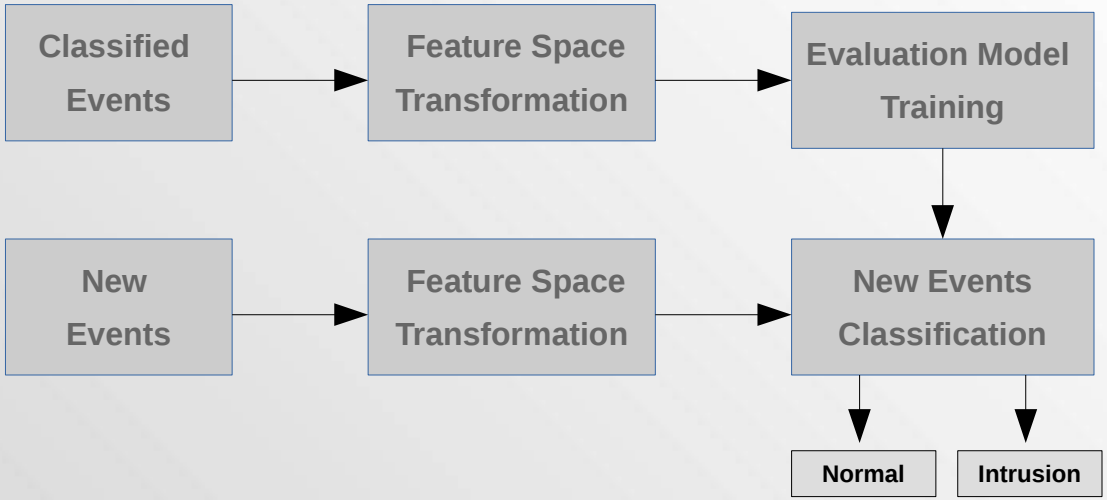
The discretized data are then extended by adding to them the **Minimum** (μ), **Maximum** (Λ), **Average** (α), and **Standard Deviation** (σ) meta-information

$$\Xi = \begin{cases} \mu = min(d_1, d_2, \ldots, d_N) \\ \Lambda = max(d_1, d_2, \ldots, d_N) \\ \alpha = \frac{1}{N} \sum_{n=1}^{N} (d_n) \\ \sigma = \sqrt{\frac{1}{N-1} \sum_{n=1}^{M} (d_n - \bar{d})^2} \end{cases}$$

Implementation

Step 3 of 3 – Event Classification

Our approach classifies the **new events** as ***normal*** or **intrusion** by using the new feature space



RESULTS

Experiments

- In addition to a canonical *k-fold cross validation* criterion, we **validated** our approach by dividing the dataset into two parts:
  - **In-sample part**, which has been used to select the internal algorithms parameters
  - **Out-of-sample part**, which has been used to evaluate the performance
- **NSL-KDD** is the real-world dataset used in order to evaluate the performance of the proposed approach
- Our competitor, experimentally selected on the basis of its performance, is **Random Forests**

Experiments

- We used the *80%* of each dataset (i.e., *PEA*, *DSA*, *RSA*, and *RAA* events) as **in-sample** data, using the remaining *20%* as **out-of-sample** data
- We evaluated our predictive model in terms of **Specificity** and **Matthews Correlation Coefficient**

Experiments

- The results show that our **DEFS** model **outperforms RF** in terms of capability to detect the intrusion in a context of datasets with a high level of unbalance, in almost all classes of events

		RF Default Model			RF DEFS Model		
Events	**DCI**	**Specificity**	**MCC**	**Average**	**Specificity**	**MCC**	**Average**
PEA	0.0015	0.5618	0.4077	0.4848	**0.7833**	**0.4835**	**0.6334**
DSA	0.6928	0.9957	0.9585	0.9771	0.9919	**0.9629**	**0.9774**
RSA	0.1826	0.9458	0.9102	0.9280	**0.9626**	**0.9175**	**0.9401**
RAA	0.0503	0.9704	0.7961	0.8833	**0.9714**	0.7891	0.8802

Conclusions

- We proposed a *Discretized Extended Feature Space* (**DEFS**) model to improve the anomaly detection performance in the IDS context
- The experimental results show its **effectiveness** in a real-world scenario
- A possible **future work** would be the experimentation of our model in the context of a classification approach based on several and different algorithms configured through an ensemble strategy

A Discretized Extended Feature Space (DEFS) Model to Improve the Anomaly Detection Performance in Network Intrusion Detection Systems

THANK YOU FOR YOUR ATTENTION

KDIR 2019, 11th International Conference on Knowledge Discovery and Information Retrieval
17÷19 September 2019 - Vienna, Austria