UDC 004.056.57

A HIERARCHICAL CORRELATION-AWARE FEATURE SELECTION METHOD FOR FLOW-BASED IDS

Hajjouz A.[1]
Scientific director – Associate Professor (Docent; qualification category: "Ordinary Docent"),
Avksenteva E.Y.[1]
[1]ITMO University
hajjouz@itmo.ru, eavksenteva@itmo.ru

Introduction. Flow-based IDS pipelines operate on telemetry spaces dominated by co-moving feature families, where collinearity entanglement inflates computational cost and destabilizes downstream attribution. This paper introduces HFS–Spearman, a label-agnostic (annotation-free) procedure for telemetry basis extraction that compresses the predictor space into a compact, dataset-specific coordinate set while preserving the intrinsic dependence structure.

Main part. HFS–Spearman induces a rank-dependence geometry by transforming Spearman correlations into a distance $d_{jk} = 1 - |\hat{\rho}_{jk}|$, then performs agglomerative structure induction (Ward linkage) to obtain a dendrogram that reveals dependence blocks in telemetry. Selection is executed through exemplar-based basis construction: each block contributes a single prototype coordinate (medoid), yielding a non-redundant representation rather than a conventional ranked feature list.

Instead of a fixed correlation threshold, the reduction is governed by a model-free, data-adaptive truncation level λ^{*} chosen via structure-preservation diagnostics (e.g., dendrogram fidelity and partition robustness indicators), turning the step into a stability-regularized compression of the predictor space.

Across heterogeneous corpora, the method yields systematic computational budget contraction (e.g., 46→23, 69→32, 67→41, 61→35, 54→33, 65→46, depending on the dataset [1-2]), with visual evidence of redundancy collapse reported for multiple datasets.

Conclusions. HFS–Spearman reframes feature selection as dependence-structure–preserving representation distillation for flow-based IDS. By combining rank-based dependence geometry, dendrogram-based block discovery, exemplar extraction, and stability-regularized truncation, it produces compact telemetry bases aligned with latency-sensitive deployment and more consistent downstream attribution behavior.

List of sources used:

1. Hajjouz A., Avksentieva E. Y. Enhancing and extending CatBoost for accurate detection and classification of DoS and DDoS attack subtypes in network traffic //Научно-технический вестник информационных технологий, механики и оптики. – 2025. – Т. 25. – №. 1. – С. 114-127.
2. Hajjouz A., Avksentieva E. Optimizing Intrusion Detection for DoS, DDoS, and Mirai Attacks Subtypes Using Hierarchical Feature Selection and CatBoost on the CICIoT2023 Dataset //Data and Metadata. – 2024. – Т. 3. – С. 577.