NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium | 978-1-6654-7716-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/NOMS56928.2023.10154455

Understanding Semantics in Feature Selection for Fault Diagnosis in Network Telemetry Data

Thomas Feltin
École polytechnique
Paris, France
thomas.feltin@polytechnique.edu

Juan Antonio Cordero Fuertes
École polytechnique
Paris, France
juan-antonio.coredro-fuertes@polytechnique.edu

Frank Brockners
Cisco Systems
Cologne, Germany
fbrockne@cisco.com

Thomas Heide Clausen
École polytechnique
Paris, France
thomas.clausen@polytechnique.edu

***Abstract*—Expert systems for fault diagnosis are computationally expensive to build and maintain, and lack scalability and inherent adaptability to unknown events or modifications in the topology of the monitored system. While data-driven feature selection mechanisms can facilitate diagnosis without the hardship of developing and maintaining expert systems, purely data-driven mechanisms lack understanding of semantic importance within a feature set, and would benefit from additional domain knowledge. Part of this additional knowledge can be extracted from meta-data. The proposed approach combines data-driven metrics and semantic information contained in the feature names to produce selections of features which best represent an underlying event. This study extends a cross entropy based optimization method to join semantic importance with data behavior. A benchmarking architecture is introduced to evaluate the benefits of semantic analysis, and demonstrate the performance and robustness of semantic feature selection on different types of faults in network telemetry datasets, modeled with the YANG data modeling language. The results illustrate the interest of such a complementary meta-data analysis for data-driven fault diagnosis, and highlight the robustness of the studied approach against variations in the input feature set.**

***Index Terms*—Fault Diagnosis, Telemetry, Feature Selection**

I. Introduction

Fault diagnosis, *i.e.*, identifying the root cause of an event, has been studied in communication networks, manufacturing, maintenance of mechanical systems, transportation, and software engineering. By mimicking processes of human reasoning, expert or rule-based systems have proven to be useful for in-depth diagnosis [1]. Most efforts of expert systems for fault diagnosis rely on the definition of a state graph, representing the known and unknown states of the system, with defined transitions, depending on the available features [2]. Typical methods include probabilistic automata and Petri nets [3], [4]. However, such fault diagnosis systems present severe scalability and adaptability issues. They require an extensive modeling stage, with full knowledge of the fault behavior, which does not scale in large, relatively complex systems. In applications such as IoT, where a variety of technologies and sensors interact with each other, or in network telemetry, where the dimension of the data changes with the network topology, telemetry data is often heterogeneous and of varying and high dimension, making it difficult to design a system which covers the entire fault behavior. Graph based expert systems also imply high computational costs in the diagnosis process when the dimension increases [5]. Being hand-crafted and domain-dependent, expert systems also lack the ability to adapt to new, unseen data [1].

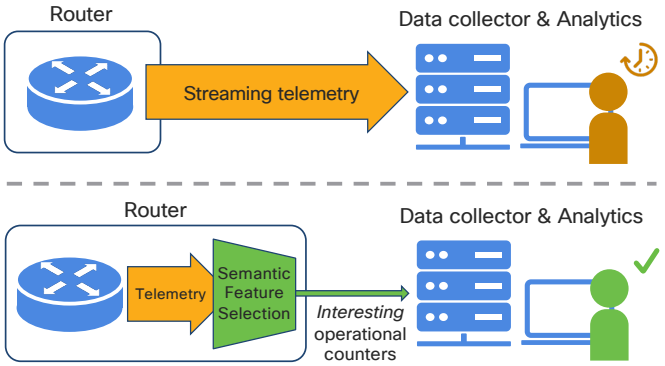


Figure 1. Semantic feature selection for fault diagnosis in network telemetry data. Instead of exporting large volumes of telemetry data to centralized servers, semantic feature selection can filter the interesting information for fault diagnosis.

In this context, robust data-driven approaches allow to (i) avoid the cost of expert systems conception and maintenance, and (ii) leverage high dimensional telemetry data to robustly diagnose events with limited domain knowledge. Insight about the inner structure of the feature set (semantics, relations, relative importance) may overcome the absence or scarceness of explicit domain knowledge. Extracting and integrating that insight about inner structure and relationships within the feature set is thus a major challenge for improving the performance of fault diagnosis systems.

One way to provide a data-driven diagnosis is distilling a set of features that are of operational importance. Selecting original features can be a simple way to assist fault diagnosis

while avoiding the modeling stage. Instead of presenting an expert with high dimensional data, feature selection narrows down the scope of investigation, as shown in Figure 1, based on a given metric, *e.g.*, individual amplitude of change, variance, or difference from a standard value.

A. Related work

Typical feature selection methods [6] identify the most important features in a dataset without the objective of explanation or fault diagnosis. Feature selection is usually intended for dimension reduction. Classical dimension reduction methods such as Principal Component Analysis, Linear Discriminant Analysis, or t-Distributed Stochastic Neighbor Embedding [7]–[9] produce artificial dimensions, *i.e.*, dimensions which are combinations (linear or not) of the original features. For the purpose of interpretability in fault diagnosis, the returned features need to correspond to original dimensions. Efforts in feature selection of original dimensions include wrapper and filter methods. Wrapper methods [10] select the original features with regards to a given clustering algorithm. Wrapper methods have proven to be computationally expensive when the original data is of high dimension [6]. Filter methods rely on data collections to define a relevance score, often based on metrics derived from entropy [11]–[13], or statistical dependencies [14] to capture the amount of information contained in the selected features. The previous selection methods have a different objective than this study, which intends to select original features for explanation instead of computational efficiency.

Feature selection for fault diagnosis has been studied to detect faults in mechanical systems and modern process industries [15]. These applications consider small input dimensions compared to this study. Other selection mechanisms for diagnosis include explanation methods in Deep Learning applications. The *explanation* process aims at selecting the original features responsible for a classification decision. Computation of the Shapley values [16] or LIME [17] score the contributions of each input feature to a classifier, in order to better understand a decision process. Although the process of selecting original features for fault diagnosis is similar, the purpose of explainable AI methods is to describe the reason for a given classification (that is, for their own decisions), whereas the objective of this study is to describe the underlying data itself. A first specification and preliminary results of this method for fault diagnosis was presented in [18].

B. Contribution

This study presents a broader conceptualization of the previous semantic analysis.

This paper presents an evaluation of the semantic feature selection method presented in [18] on network telemetry datasets, a hybrid selection process which combines data-driven metrics and semantic analysis of meta-data. This approach produces a representation for network fault events, extracted from the telemetry available, that can be used for fault diagnosis. This study presents a broader conceptualization of the previous semantic analysis. The added contributions of this paper are the following:

- The demonstration that data-driven feature selection methods fail to identify semantic feature importance relationships.
- This introduction of a novel benchmark for evaluating the performance and robustness of selection methods for event diagnosis on telemetry data.
- The presentation of benchmarking results, for both data-driven and semantic feature selection methods, on data retrieved from routers running the Cisco IOS-XR operating system, modeled with the YANG data modeling language [19]. This benchmark demonstrates both the performance and robustness of the semantic feature selection method for fault diagnosis.

C. Paper outline

The remainder of this paper is organized as follows: section II describes the problem space and specificities of network telemetry data. section III presents the limitations of data-driven selection methods and highlights the need for additional information. section IV presents the meta-data based importance estimation method and its inclusion in the selection process. Finally, section V presents the experiment setup and results for performance and robustness evaluation on network telemetry datasets.

II. Data description and preprocessing

This section presents the data particularities of telemetry datasets and preprocessing mechanisms in the context of fault diagnosis.

A. Telemetry data properties

Telemetry data takes the form of a time-series with each feature representing the value of one particular sensor in a system over time. Specifically in the context of network telemetry the following data properties can be described.

1) High, variable dimensionality: In telemetry applications where the cost of an individual sensor, or of measuring an individual feature, is relatively low (as it is the case in networking, software engineering, IoT), telemetry datasets may be high-dimensional. In dynamic systems, dimensionality itself may change over time, as features appear or disappear. In particular, in network telemetry, performance of a network interface i may be described through n_i features or dimensions including interface i's byte count, data rate, queue occupation, etc. Enabling or disabling interface i in a network thus leads to an increase or decrease of the dimensionality of corresponding network telemetry dataset by n_i. Dealing with dataset *holes* that result from such variability may require additional data preprocessing.

2) Heterogeneity: The data values can be of different data types and formats, *e.g.*, in the network telemetry datasets used in this study, on an individual router, the numerical features can be positive incremental integer values ranging in the billions, *e.g.*, byte counts, or non-monotonic functions

ranging from 0 to 1, *e.g.*, CPU consumption. Comparing data of different nature also requires additional preprocessing.

3) Aggregation level: Telemetry datasets are also heterogeneous in the aggregation level of each feature. Telemetry applications usually monitor complex systems, composed of different subsystems or services, such as network routers or mechanical systems composed of individual components and services. Telemetry data is usually composed of (i) features describing the state of a single element (referred as *individual* feature in this paper, *e.g.*, a single byte counter), and (ii) features which are aggregations of several sources of information (referred as *compound* features in this paper, *e.g.*, the total number of open connections on a router).

B. Telemetry data preprocessing

The properties presented in section II-A highlight the challenges of data-driven fault diagnosis. The high and varying dimensionality cause dimensions to appear and disappear from the feature set dynamically, without indication on their signification or relevance. Heterogeneity complicates the use of most data-driven approaches which assume the identical nature of all features. Different aggregation levels implies different levels of importance in the feature set, which only stems from domain knowledge. However, with the assumption that the data around an event is fixed, it is possible to define a window around the event time where the data can be relieved of some of these properties. For example, differentiating the incremental features and performing min-max scaling can solve the heterogeneity problem. This differentiation can be performed in real-time by estimating which features are incremental during a bootstrapping period, the duration of this period being considered long enough to simply estimate that any monotonously increasing features is incremental. The dynamic dimension problem can be handled by padding missing values with zeros, which narrows down the dataset to a time-series of fixed dimension for further processing.

This preprocessing is not optimal: (i) min-max scaling does not handle unbalanced data well, (ii) the bootstrapping period can be too short and consider non-incremental data as incremental, and (iii) zero-padding dynamic dimensions can create artificial abrupt changes in the data which alter detection and diagnosis. Further optimization is outside of the scope of this paper since the presented preprocessing mechanism has proven to provide reasonable results in section V.

III. Data-driven methods

This section presents data-driven selection methods for fault diagnosis. These methods take an event time, t_0, as a given input[1], and return a ranking of the features with weights representing how much each feature changes within a window around the event time $[t_0 - w, t_0 + w]$. Throughout this section, the pre-processed time-series data is annotated as $\mathbf{S} = \{s_{n,t}\}_{n \leq N, t \leq T}$ with $s_{n,t}$ being the value of feature n at time t.

[1]Event detection itself and mechanisms for determining the time $t_0 = t(e)$ of an event e are outside the scope of this paper; a change-point detection method [20] could be used for this purpose.

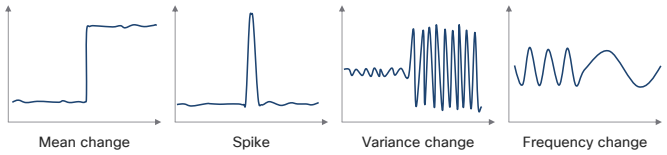


Figure 2. Change shapes present in telemetry datasets. For simplicity, the univariate change amplitude metric only considers mean and spike changes.

Three approaches are tested on network telemetry datasets to cover a range of data-driven change amplitude metrics: the univariate change amplitude method, which computes the change amplitude for every feature independently by looking at univariate time-series data (section III-A), a linear multivariate change amplitude method, which considers N-dimensional data points for every time step (section III-B), and a non-linear N-dimensional method which identifies feature contributions to a non-linear classifier (section III-C). A comparative analysis of the examined methods concludes the section (section III-D).

A. Univariate Change Amplitude (UCA)

The univariate change amplitude is defined, for every feature, as the maximum value between the normalized mean value change between windows respectively before and after the time of the event, and the normalized spike amplitude at the time of the event. With $\mathbb{E}_{t\in[t_i:t_f]}(s_{n,t})$ being the temporal-mean value of the normalized data between times t_i and t_f, and A the value of the spike amplitude on a window around the time of the event $[t_0 - \epsilon, t_0 + \epsilon]$, for every feature n, the univariate change amplitude is expressed as:

$$\sigma(\mathbf{s}_n, t_0) = \max\Big(\big|\mathbb{E}_{t\in[t_0-w:t_0]}(s_{n,t}) - \mathbb{E}_{t\in[t_0:t_0+w]}(s_{n,t})\big| \\ \Big|A - \frac{1}{2}(\mathbb{E}_{t\in[t_0-w:t_0-\epsilon]}(s_{n,t}) + \mathbb{E}_{t\in[t_0+\epsilon:t_0+w]}(s_{n,t})\Big|\Big) \tag{1}$$

The univariate change amplitude metric scores every feature independently by looking only at the univariate data in a given time frame. Given this data, when an event occurs in the system, the observed change can show several patterns, as shown in Figure 2, *e.g.*, mean value change, variance change, spikes, frequency change. The metric needs to provide comparable values for different shapes or data type, *e.g.*, numerical, categorical, incremental or not. Limiting this metric to first moment statistics, *i.e.*, changes in mean value or spikes, simplifies the metric which can be used as a baseline for comparison with the more elaborate methods below.

B. Linear Discriminant Analysis (LDA)

The approach taken in this section considers N-dimensional data points instead of univariate time-series. The method used is the Linear Discriminant Analysis (LDA) [8], which computes the coordinates of the hyperplane which maximizes

separability between N-dimensional data points before and after the event. With $\mathbf{s}_t$ defined as the N-dimensional vector of feature values at time t, the weights in the final ranking correspond to the coefficients of the hyperplane $\mathbf{v}^*$, defined as:

$$\mathbf{v}^* = (\mathbf{\Sigma}_- + \mathbf{\Sigma}_+)^{-1}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$$
$$\text{with} \qquad \boldsymbol{\mu}_\pm = \mathbb{E}_{t\in[t_0:t_0\pm w]}(\mathbf{s}_t)$$
$$\text{and} \qquad \mathbf{\Sigma}_\pm = \sum_{t\in[t_0:t_0\pm w]} (\mathbf{s}_t - \boldsymbol{\mu}_\pm)(\mathbf{s}_t - \boldsymbol{\mu}_\pm)^\top \tag{2}$$

With this classification between points in windows around the event, the ranking will be less sensitive to outliers and represent general tendencies in the data. However, the classifier will be unable to discriminate between classes if the discriminatory, *i.e.*, this method makes assumptions on the distribution of the measurements being Gaussian within both classes.

C. Non-linear classifier

The chosen non-linear classifier is a random forest trained to classify points before and after the event time, and the feature contributions are extracted by leveraging the SHAP methodology [16], which relies on the computation of the Shapley values.

Similarly to LDA, this ranking considers N-dimensional points at any given time t, without assuming a distribution over the two classes, which enables the consideration of non-linear relationships between points before and after the event time t_0. The SHAP methodology is originally designed for explainable AI, where methods are developed to find the original features which contribute most to a classifier decision. In order to apply this method, the binary random forest classifier is first trained to classify points before and after the event time, before applying the explanation methodology which ranks features by their contribution to the classifier predictions.

D. Limitations illustration

The three methods are applied on an example network telemetry dataset of 23650 individual features describing the state of a router running the Cisco IOS-XR operating system, when an interface shuts down[2].

- *Univariate change amplitude*: Among the highest ranked features are mostly compound features such as interface counters, Bidirectional Forwarding Detection (BFD) sessions state counters, Border Gateway Protocol (BGP) neighbour counters, as well as individual features such as the last Routing Information Base (RIB) version, or negotiated intervals.
- *Linear discriminant analysis*: The highest contributing features are traffic counters, data rates and neighbour advertisement message counters.
- *Non-linear classifier*: The highest ranked features are exclusively packet counters and data rates related to the state of the interfaces neighbouring the shutdown interface.

[2]Because of the solutions verbosity, the 50 highest ranking features for each method can be found at https://github.com/tfeltin/sefset_results/blob/master/datadriven.md

Although all the selected features are either relevant to the interface shutdown, or linked to a consequence of this event, none of the methods above place the most important compound features first, *i.e.*, features representing the number of active interfaces. This suggest that the highest ranked features are highly changing in value, but do not necessarily correspond to anything meaningful to the diagnosis.

Some features are important than others in the diagnosis process, albeit identical from a data-driven point of view. For example, the feature counting active interfaces `up-interface-count` and the feature describing the last version of the RIB table `last-rib-version` have an identical behavior around the event, *i.e.*, a step from an integer value to another, yet `up-interface-count` can be considered as the most important indicator of an interface shutting down, while `last-rib-version` describes a consequence, as a part of the re-routing mechanism. From a purely data-driven approach, the two features are indistinguishable and are both included among the highest ranked features.

From this analysis, the conclusion is drawn that the described data-driven approaches cannot capture importance relationships between features, and additional information needs to be taken into account to identify important features in telemetry datasets.

IV. Semantic feature selection

This section presents a generalized approach to the semantic importance estimation presented in [18] which uses meta-data information contained in feature names.

A. Estimating semantic importance

Not all features in a telemetry dataset have the same relevance for the diagnosis of a given fault. For example, compound features usually are more relevant because they offer an aggregated perspective of the state of the system, compared to individual features which only describe a single component. Section III having illustrated that such relevance cannot be retrieved exclusively from data, this paper suggests to estimate the relative importance of features for fault diagnosis by exploiting available meta-data, which carries semantic information. The applied methodology extends the feature name based semantic importance estimation of [18].

Term Frequency-Inverse Document Frequency (TF-IDF) [21] is a measure of the importance of a word (or term) to a document in a corpus of documents; word *importance* is quantified as the relative frequency of the word in the document (term frequency), divided over an estimation of the information provided by the word in the whole corpus of documents (inverse document frequency).

$$\text{tfidf} = \text{tf} \cdot \log\left(N/\text{df}\right) \tag{3}$$

In our case, documents are the feature names, and words are the *tokens* forming a feature name; these can be individual

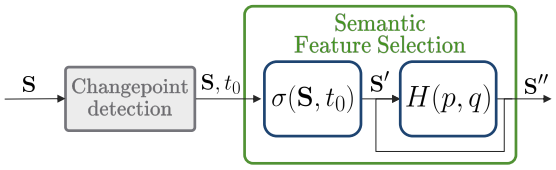


Figure 3. Flow chart of semantic feature selection for fault diagnosis. It is assumed that the diagnosis process is associated with a change-point detector (outside of the scope of this work), which produces time stamps t_0 of detected events from streaming telemetry data $\mathbf{S}$. The semantic feature selection uses a data-driven change amplitude metric $\sigma(\mathbf{S}, t_0)$ to define the optimization objective based on cross entropy $H(p, q)$ in Equation 6.

words or groups of words that go together in a feature name. For example, the feature name `up-interface-count` consists of three tokens (individual words): `up`, `interface`, and `count`.

In Equation 3, tf is thus the number of times the token appears in the feature name, df the number of feature names containing the token, and N the dimension of the dataset. The importance of an entire feature name, *i.e.*, the semantic importance of one feature in the dataset, is estimated as the average importance of its tokens.

The importance estimation was computed for the same network telemetry dataset used in section III-D. Within the top 20 highest scoring features are interface counters, CPU and memory utilization counters, BFD session counters, and ICMP TTL exceeded counters. It can be observed that the highest ranking features are compound features *e.g.*, containing the words `summary` or `total` in their feature name. This confirms the assumption that compound features can be estimated to be semantically important.

Feature names are usually short and contain very few token repetitions. It can be estimated that the number of times a token appears in a feature name $tf \approx 1$, in which case the importance of a token can simply be approximated by the inverse frequency of its occurrence in the dataset. This observation shows that this metric is relying on the *discriminative* power of a feature name. The more unique a feature is in a dataset, the more information it contains when impacted by an event. Token importance is defined as a distribution over $\mathcal{T}$, the space of all existing tokens in the feature set. Considering $\tau \in \mathcal{T}$ a token, the importance of this token in the set is approximated as p:

$$p(\tau) = n(\tau)/T \tag{4}$$

with $n(\tau)$ the number of times the token τ appears in the set, and T the sum of the number of tokens in each feature name in the set.

B. Selection cross entropy

Assuming a subset of features with a token distribution q is taken from the original feature set with token distribution p. To describe an event, this subset is expected to preserve tokens from the original set which carry the most information, *i.e.*, rare token instances in distribution p. This is quantified through a measure of cross entropy $H(p, q)$ between the initial and final token distributions p and q, defined as:

$$H(p, q) = -\sum_{\tau \in \mathcal{T}} p(\tau) \log q(\tau) \tag{5}$$

A high cross entropy implies a low likelihood is related to semantic quality, *i.e.*, the ability for an operator to understand the selection. The advantage of using this metric is twofold: *(i)* it favors the selection of features composed of rare tokens, identified as carrying the most information, and *(ii)* it favors specificity in the distribution, which favors feature selections with a small number of tokens. The selection which maximizes the cross entropy value is just the one feature which contains the rarest tokens in the set. Having a single feature as selection is trivial to interpret, even more so when its tokens preserve the most information from the original set. However, in the context of fault diagnosis, this needs to be joined with feature contributions to the event, which is why cross entropy is jointly maximized with a change amplitude metric presented in section III.

C. Optimization process

The semantic feature selection is defined as an optimization process which jointly maximizes the data-driven change metric and the cross entropy based semantic importance estimation, as shown in Figure 3. The final optimization objective is defined as [18]:

$$\mathcal{L}_\alpha(\mathbf{S}, t_0, p, q) = (1 - e^{-\frac{|\mathbf{S}|}{\alpha}}) H(p, q) \frac{1}{|\mathbf{S}|} \sum_{\mathbf{s} \in \mathbf{S}} \sigma(\mathbf{s}, t_0) \tag{6}$$

where $\sigma(\mathbf{s}, t_0)$ is the univariate change amplitude score from section III-A, and α is a regularization parameter, which aims at penalizing very small selections and can be used as a tuning parameter for the size of the selection.

The optimization is defined a greedy process. The input variables are the event time t_0 and the multivariate time-series data $\mathbf{S}$. The change score $\sigma(\mathbf{s}, t_0)$ is computed for each univariate time series $\mathbf{s} \in \mathbf{S}$ in the full dataset and the scores are ranked in descending order. The initializing step selects the N_i features with highest change scores as the initial selection, with N_i defined depending on the use-case ($N_i = 500$ in the network telemetry application used in section V). At every step, the selection mechanism performs two operations. First, for every feature that is not in the selection, the score is computed for the selection with this extra feature included. Every feature which improves the score of the current selection when added is appended to the current selection. Then, the same operation is performed, by instead trying to remove each feature in the selection. Every removed feature which improves the score is removed from the current selection. The optimization process stops when no addition or removal improves the score. If this criterion is not reached after I_{max} iterations, the process stops and returns the current selection with the notification that an optimum was not reached.

D. Illustration

The output of the selection method is a list of original input features which are deemed to best describe the underlying event. For example in the network telemetry datasets in this study, for a routing loop, data was collected on one of the devices through which the traffic is looping. With feature names modeled with YANG and more than 40 thousand input features, the output of semantic feature selection on a routing loop is the following 6 features:

```
- Cisco-IOS-XR-ipv4-io-oper:ipv4-network/
nodes/node/statistics/traffic[node-name=
0/1/CPU0]output
- Cisco-IOS-XR-ipv4-io-oper:ipv4-network/
nodes/node/statistics/traffic[node-name=
0/1/CPU0]hopcount-sent
- Cisco-IOS-XR-infra-statsd-oper:infra-
statistics/interfaces/interface/latest/
protocols/protocol[interface-name=
HundredGigE0/1/0/34 protocol-name=
IPV4_UNICAST]bytes-sent
- Cisco-IOS-XR-infra-statsd-oper:infra-
statistics/interfaces/interface/latest/
interfaces-mib-counters[interface-name=
HundredGigE0/1/0/34]bytes-sent
- Cisco-IOS-XR-infra-statsd-oper:infra-
statistics/interfaces/interface/latest/
generic-counters[interface-name=
HundredGigE0/1/0/34]bytes-sent
- Cisco-IOS-XR-pfi-im-cmd-oper:interfaces/
interface-xr/interface[interface-name=
HundredGigE0/1/0/34]bytes-sent
```

The first two feature in this list are the ICMP output and hop count exceeded features, followed by traffic counters on the interface which is seeing the incoming traffic from the loop.

V. Benchmark

Comparing this feature selection for fault diagnosis method to existing solutions is difficult because the literature lacks a dedicated evaluation. This section introduces a benchmark to address this issue and identify the added benefit of a semantic analysis of meta-data for fault diagnosis. This approach also aims to evaluate the robustness of the method with variations in the collected input features, in order to quantify the dependency of the semantic method on the input feature set.

A. Datasets

The datasets used in this benchmark are extracted from several devices in a Clos-topology lab environment with simulated traffic and inserted events[3]. Each dataset contains 20 to 40 thousand individual features depending on the device. The datasets in the benchmark each contain one single inserted event. The time of the event is known to the selection mechanism. The purpose is to evaluate every selection output in an isolated way.

Four injected events are used in this benchmark and for each event, data was collected from different devices : (i) interface admin shutdown, on the two devices at the ends of the disconnected link, (ii) BFD failure (filtering BFD message to trigger a failure), (iii) black hole (removing FIB entries to cause silent packet drops), collected on the concerned device, and (iv) a routing loop (by adding static routes), with data collected from the three devices concerned by the loop.

B. Metrics for feature selection

Objectively evaluating a feature selection for a diagnosis process is complicated since the feature importance during diagnosis is subjective and operator dependent. With this observation, defining a metric which quantifies precisely how far the output is from an ideal is impossible. Although the precise ground truth is intractable, the metrics defined in this section aim to get an assessment of how the method is performing.

The idea is to define an upper and lower bound on what the acceptable outputs are for this method, based on the described event. This paper defines the following two metrics that will act as such:

- *Precision*: defined as the proportion of selected features in the output selection that are related to the input event. Among all features, those related to an event are defined as a pre-established subset which are deemed relevant to select when an even occurs. This metric quantifies the relevancy of the selection with regards to the input event.
- *Completeness*: Within the list of all acceptable features to describe an input event, a subset of these features are essential to diagnosis. Completeness quantifies whether the output contains this very minimal requirement, *i.e.*, completeness is the proportion of this minimal subset which is present in the output.

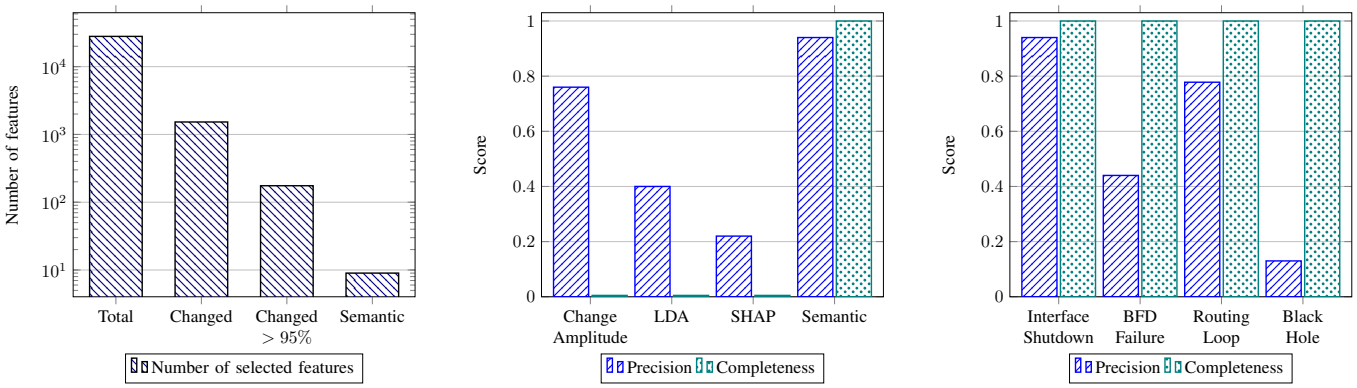
Along with performance, this paper defines one other metric to evaluate the robustness of the method. As the selection method highly depends on the input data, over-fitting to either the raw data or the feature set is a possibility. Therefore, the following metric is used:

- *Robustness*: quantifies the degree of variation in the output when the input collection (feature set) is modified. Intuitively, since the importance estimation highly depends on the probability distributions defined in section IV-B, it is expected that variations in the feature set may cause significant variations in the selection. The robustness evaluation aims at identifying precisely which type of variation in the feature set is associated with which degree of variation in the final selection.

C. Ground truth definition for evaluation

For each event type, computing the performance and robustness metrics depends on the definition of a set of features which act as ground truth for both precision and completeness.

[3]Details on the topology in https://github.com/cisco-ie/telemetry



(a) Comparison of number of reduced dimensions

(b) Precision and Completeness scores of methods for an interface shutdown

(c) Precision and Completeness scores

Figure 4. Comparative benchmarking results between semantic and data-driven feature selection methods for event diagnosis in telemetry data. Figure 4(a) shows the amount of dimensions reduced, comparing the full feature set (Total), features with changing data (Changed), features with value changes of more than 95% (Changed $> 95\%$), and features selected with semantic analysis (Semantic). Figure 4(b) shows the benchmarking scores of the data-driven methods described in section III and compares them to semantic feature selection on an interface shutdown. Figure 4(c) shows the benchmarking results of semantic feature selection on the four events in the benchmark.

Event type	**Regular expressions**
Interface Shutdown	**P :** `interface.*summary` / `HundredGigE0/0/0/N` **C :** `interface-count`
BFD Failure	**P :** `bfd` **C :** `bfd.*summary.*count`
Routing Loop	**P :** `icmp` / `bgp` / `hopcount` / `bytes-(sent|received)` / `ipv4-io-oper.*traffic` **C :** `hopcount`
Black Hole	**P :** `rate` / `load` / `icmp` / `ipv4-io-oper.*traffic` / `connections-(established|accepted|closed)` **C :** `unreachable-received`

Figure 5. Ground truth definition for the four event types contained in the benchmark, for network telemetry data modeled with YANG. Regular expressions are defined for precision (P) and completeness (C).

The precision ground truth feature set should contain all features which are linked to the event, and the completeness ground truth feature set should contain the minimal set of features expected from the selection. For simplicity in this network telemetry use-case, the two feature sets are defined as a collection of regular expressions. This benchmark defines a ground truth for four event types in the context of network telemetry: interface shutdowns, routing loops, BFD session failures, and black holes. Figure 5 shows the ground truth definition for datasets using Cisco YANG models.

D. Robustness evaluation

The space of all input subscriptions is too large to be fully enumerated. The robustness evaluations presented in this study target two scenarios: (i) incomplete feature sets, *e.g.*, caused by collection errors, and (ii) variations in token distributions, *e.g.*, caused by different subscriptions or operators.

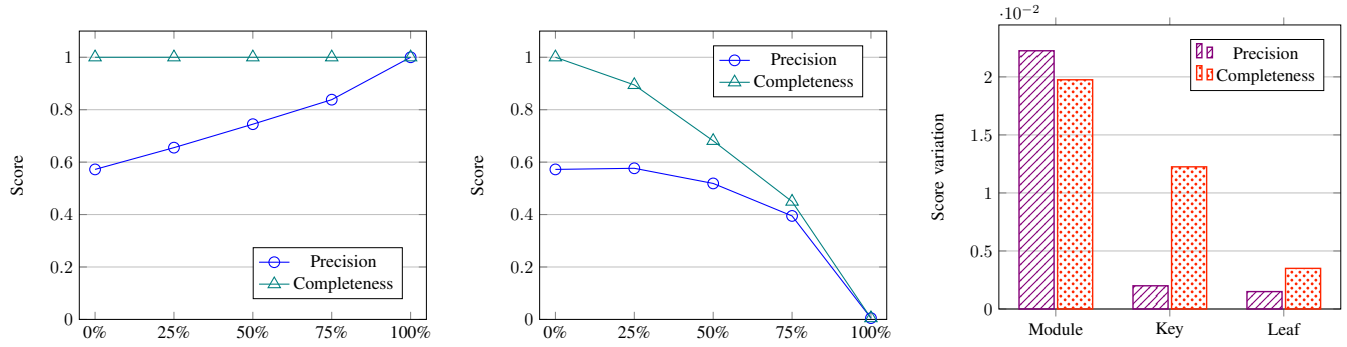
1) Random removal of entire modules: From the list of subscriptions contained in the datasets, this method measures the impact of removing all the data contained in one or more modules on the selection output. Different quantities of removal are tested, *i.e.*, 25%, 50%, and 75% of modules. Two metrics are extracted from this evaluation based on whether the removed features belong to the ground truth defined in section V-C or not: sensitivity, *i.e.*, score variations when ground truth related features are removed, and consistency, *i.e.*, score variations when non ground truth features are removed.

2) Changing feature name distribution: This methodology willingly makes frequent tokens rarer, as frequency is the determining factor in the importance estimation. Since YANG paths can be split into three types of tokens, *i.e.*, modules, keys, and leaves [19], this process is done independently for each token. The top 5%, 10%, 25%, 50% most frequent tokens in each type are made rare, by keeping only one feature among all the ones containing the given token. Robustness scores are computed for precision and completeness for each token type, for a more granular view. The robustness scores are the average score variations caused by altering one token frequency.

E. Results and discussion

Figure 4 shows the number of selected features and their performance evaluation results on the benchmark for the four previous event types, and compares them with data-driven methods. The results vary depending on the event, with perfect completeness scores and varying precision[4]. Good completeness with lower values in precision such as is the case for the black hole evaluated in the benchmark corresponds to

[4]The complete selection results can be found at https://github.com/tfeltin/sefset_results/blob/master/results.md



(a) Consistency to removal of modules not contained in the ground truth

(b) Sensitivity to removal of modules contained in the ground truth

(c) Error induced by altered token distributions.

Figure 6. Robustness evaluation. Figure 6(a) (resp. Figure 6(b)) shows the evolution of the precision and completeness of selections when removing modules which are not included in the ground truth (resp. which are included in the ground truth). Figure 6(c) shows the impact of altering token distributions for three token types in YANG, as described in section V-D2.

verbose selections with parasite information, but still the right essential indicators of the event. Overall, the semantic analysis identifies the 10 to 20 counters which best represent an event in datasets of 20 to 40 thousand individual counters. Additionally, the comparison with a data-driven method shows the effect of semantic analysis. Although the data-driven methods find features which are linked to the event, they are unable to find the most important feature which are related to an event, as hinted in section II, and confirmed by the completeness score.

Figure 6 shows the results of the robustness evaluation. Figure 6(a) and 6(b) show the robustness evaluation scores from the first method, *i.e.*, the achieved variations in precision and completeness when removing features from the input set. As disclosed by the first robustness test, when removing 25% to 50% of the input features, the selections are still relevant. Additionally, with around 60 modules in the studied feature sets, when dividing these results by the number of removed modules, it can be argued that removing a single module has very little impact overall.

Figure 6(c) shows the robustness evaluation using the second method, *i.e.*, willingly altering the distribution of tokens in the feature set. The main takeaway is that the impact of drastically changing a token's frequency is almost insignificant on the final selection, which is unexpected. Changing the frequency should mean changing its importance estimation, and break the logic of the semantic analysis, yet the selections stay similar. One potential cause is the intrinsic logic that exists in the combinations between the different tokens in the naming model. When removing all the features that contain one given leaf name, the modules and key values consequently removed from the dataset are not randomly selected. This robustness evaluation shows that making the most frequent tokens rare does not impact the importance estimation severely enough to invalidate the semantic analysis.

VI. Conclusion

Telemetry data often being heterogeneous, and of high, varying dimensionality, traditional expert systems lack adaptability and require extensive design and maintenance efforts. Data-driven approaches can be studied as light-weight and robust solutions. In that regard, semantic feature selection for diagnosis is a general solution offering a hint as to which original features best represent the event.

This paper has shown that purely data-driven feature selection methods for fault diagnosis are inefficient, and unable to identify the most important features to describe an event. Although such importance relationships are defined by domain knowledge, this paper studies an approach for estimating these semantic importance relationships by studying the meta-data information contained in available feature names. This semantic analysis produces more complete and precise selections, while significantly reducing the number of features to analyze.

With the elaboration of a benchmark for evaluating feature selection methods for fault diagnosis, this study had shown the performance and robustness of a semantic approach on network telemetry datasets, with feature names derived from associated YANG models. The benchmarking results show the added improvement of semantic analysis compared to data-driven methods, with an average precision score 1.5x higher than data-driven methods and a significant completeness score improvement (0% for purely data-driven methods). Additionally, this study has evaluated the method robustness to be under variations on average in the output scores for removed or altered modules. This result shows a robustness of the semantic analysis against strong variations in the input feature set, indicating the ability of the method to capture semantic relationship between features independently of the input feature set. This study has shown experimental results on four event types, for telemetry datasets using the YANG modeling language. A potential extension of this work could include testing with a higher variety of fault types, and modeling structures (*e.g.*, SMI, OpenConfig).

References

[1] C. Angeli *et al.*, "Diagnostic expert systems: From expert's knowledge to real-time systems," *Advanced knowledge based systems: Model, applications & research*, vol. 1, pp. 50–73, 2010.

[2] D. Wang, X. Wang, and Z. Li, "State-based fault diagnosis of discrete-event systems with partially observable outputs," *Information Sciences*, vol. 529, pp. 87 – 100, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025520303327

[3] C. G. Cassandras and S. Lafortune, *Introduction to discrete event systems*. Springer Science & Business Media, 2009.

[4] C. Seatzu, M. Silva, and J. H. Van Schuppen, *Control of discrete-event systems*. Springer, 2013, vol. 433.

[5] J. Zaytoon and S. Lafortune, "Overview of fault diagnosis methods for discrete event systems," *Annual Reviews in Control*, vol. 37, no. 2, pp. 308 – 320, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1367578813000552

[6] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, 2019. [Online]. Available: https://doi.org/10.1007/s10462-019-09682-y

[7] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987, proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0169743987800849

[8] K. Fukuaga, "Introduction to statistical pattern classification," *Patt. Recognit*, vol. 30, no. 7, pp. 1145–1149, 1990.

[9] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[10] J. Yao, Q. Mao, S. Goodison, V. Mai, and Y. Sun, "Feature selection for unsupervised learning through local learning," *Pattern Recognition Letters*, vol. 53, pp. 100 – 107, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865514003559

[11] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, Nov 1997, pp. 532–539.

[12] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel Unsupervised Feature Filtering of Biological Data," *Bioinformatics*, vol. 22, no. 14, pp. e507–e513, 07 2006. [Online]. Available: https://doi.org/10.1093/bioinformatics/btl214

[13] V. M. Rao and V. N. Sastry, "Unsupervised feature ranking based on representation entropy," in *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*, March 2012, pp. 421–425.

[14] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, March 2002.

[15] T. W. Rauber, F. de Assis Boldt, and F. M. Varejão, "Heterogeneous feature models and feature selection applied to bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 637–646, 2015.

[16] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938

[18] T. Feltin, P. Foroughi, W. Shao, F. Brockners, and T. H. Clausen, "Semantic feature selection for network telemetry event description," in *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2020, pp. 1–6.

[19] E. M. Bjorklund, "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)," Internet Requests for Comments, RFC Editor, RFC 6020, October 2010. [Online]. Available: https://tools.ietf.org/html/rfc6020

[20] P. Foroughi, W. Shao, F. Brockners, and J.-L. Rougier, "DESTIN: Detecting state transitions in network elements," in *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, 2021, pp. 161–169.

[21] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, vol. 39, no. 1, pp. 45–65, 2003.