Network Feature Selection based on Machine Learning for Resource Management

Ons Aouedi, Kandaraj Piamrat, Benoît Parrein

Network Feature Selection based on Machine Learning for Resource Management

Ons AOUEDI, Kandaraj PIAMRAT and Benoît PARREIN

University of Nantes, LS2N (UMR 6004)

{firstname.lastname}@ls2n.fr

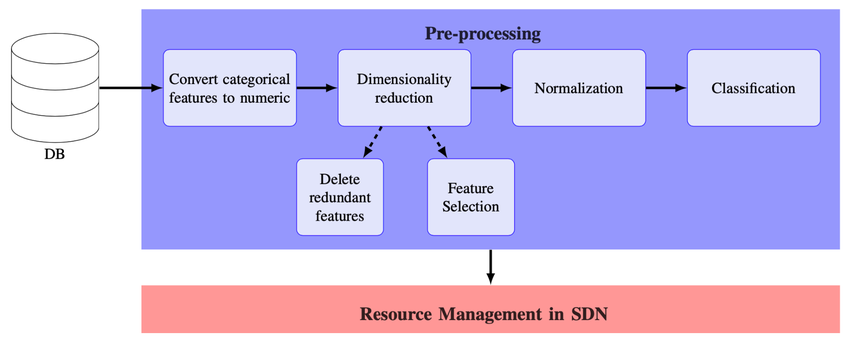
Motivation

Resource management in SDN (e.g. network slicing) is an emerging area that attracts the attention of academia and industry. It is an indispensable technology in 5G systems. To effectively manage and optimize network resources, more intelligence needs to be deployed. Therefore, combining real network data and Machine Learning (ML) with the benefits of SDN can be a promising solution to manage the network resources in an automated and intelligent way. However, a real network dataset can have redundant and unneeded features. Also, ML algorithms are as good as the quality of data and the SDN is a time-critical system that requires real-time processing and decision. Thus, data preprocessing is a necessary task, which helps to keep the relevant features and makes the prediction quicker and more accurate.

This work presents a comparative analysis between two feature selection methods, which are **Recursive Feature Elimination (RFE)** and **Information Gain Attribute Evaluation (InfoGain)**, using several classifiers on different reduced versions of the network's dataset.

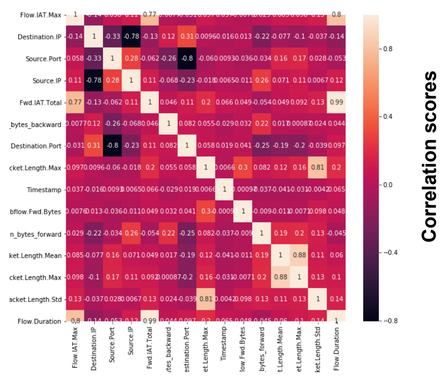
Proposition

To prepare our data for resource management in SDN, methods of attribute selection (RFE and InfoGain) followed by the classification have been used to find a subset of appropriate features from our dataset.



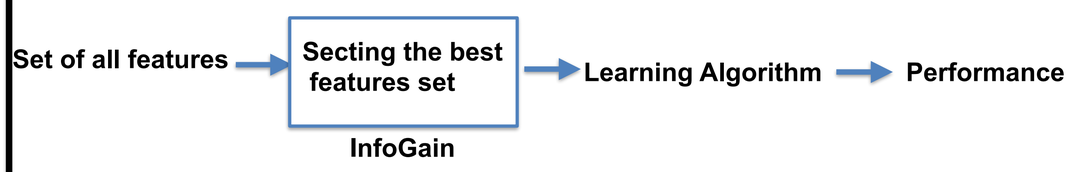
Steps of our proposition

1. Correlation matrix states how the features are correlated to each other and with target variables

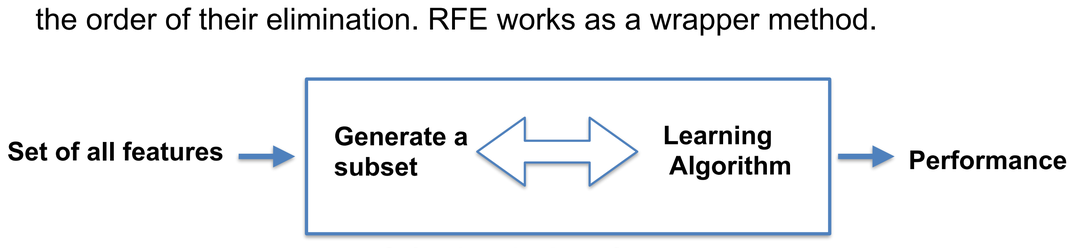


2. Feature selection

- **InfoGain [1]**: is based on the calculation of the entropy and measures the mutual information provided by X (features) on Y (target). It works as a filter method.



- **RFE [2]**: its goal is to select features by recursively considering smaller and smaller sets of features. It repeatedly creates models and keeps aside the best performing features at each iteration. Then it ranks the features based on the order of their elimination. RFE works as a wrapper method.



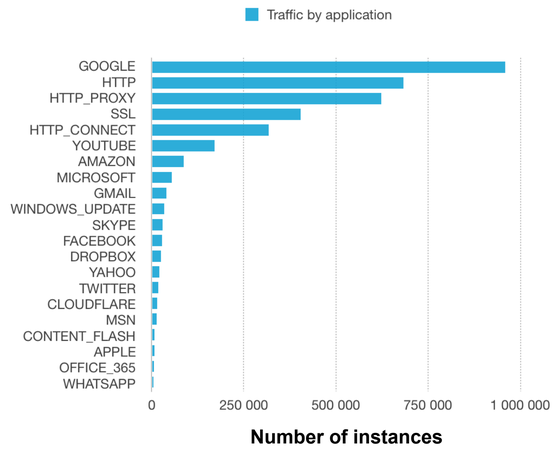
3. Classification is a data mining task that takes some types of input data and maps it to a discrete label. To classify we used **Decision tree**, **Random Forest**, **XGBoost** and **AdaBoost**.

Dataset description

A real network dataset has been used:

- Collected in network section from Universidad Del Cauca, Popay an, Colombia [3].
- 87 variables and 3,577,296 instances.
- Label (Google, WhatsApp, Netflix etc).

80% of the dataset was used for training and 20% for testing.



Results

CLASSIFICATION ACCURACY WITH RFE ON DIFFERENT SUBSET OF FEATURES

Classifiers	Top 5	Top 10	Top 15	Top 20	Top 25
Decision Tree	70.40 %	79.87 %	81.90 %	81.88 %	81.55 %
Random Forest	75.53 %	84.29 %	85.17 %	85.08 %	84.66 %
XGBoost	75.36 %	86.33 %	89.01 %	88.89 %	88.86 %
AdaBoost	75.32 %	86.65 %	87.42 %	86.99 %	85.85 %

CLASSIFICATION ACCURACY WITH InfoGain ON DIFFERENT SUBSET OF FEATURES-1

Classifiers	Top 5	Top 10	Top 15	Top 20	Top 25
Decision Tree	76.52 %	81.38 %	81.15 %	80.65 %	80.55 %
Random Forest	77.23 %	85.14 %	84.93 %	84.62 %	84.13 %
XGBoost	77.67 %	87.50 %	87.00 %	87.13 %	86.97 %
AdaBoost	72.23 %	86.22 %	86.07 %	85.63 %	85.36 %

These features are (for top 15)

Flow.IAT.Max	Init_Win_Bytes_Backward	Init_Win_Bytes_Forward
DestinationIP	DestinationPort	Bwd.Packet.Length.Mean
SourcePort	Fwd.Packet.length.Max	Bwd.Packet.Length.Max
SourceIP	Timestamp	Fwd.Packet.Length.Std
Fwd.IAT.Total	Subflow.Fwd.Bytes	Flow.Duration

Conclusion

- A comparative analysis between InfoGain and RFE
- The performance does not increase with more feature sets
- Top 15 features selected by RFE maximize classifiers accuracies
- These features can be an appropriate subset to characterize our class (Applications)

References

[1] J. Novakovic, Using Information Gain Attribute Evaluation to Classify Sonar Targets,*The 17th Telecommunication forum TELFOR*, 2009.

[2] Kohavi, R. & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97:12, 273–324.

[3] J.S. Rojas, A. Rendon, J. Corral, Personalized Service Degradation Policieson OTT Applications Based on the Consumption Behavior of Users, International Conference on Computational Science and Its Applications, 2018 pp. 543–557.