Anomaly detection for the Fink broker

Intern
Pierre Cavalier

Supervisor

Julien Peloton

IJCLab Irène Joliot-Curie

université PARIS-SACLAY

20th July 2023

Plan

1. Introduction
    - What is Fink ?
    - Dataset
2. Goal and motivation
3. Graphs
    - Definition
    - Creation and performance
4. Anomalous and ambiguous data
5. Perspectives

Introduction

What is Fink ?

- Interface between telescopes and users

Introduction

What is Fink ?

- Interface between telescopes and users
- Community project since 2019

Introduction

What is Fink ?

- Interface between telescopes and users
- Community project since 2019
- $\sim$ 50 members in 13 countries

FINK

The data set

A dataset element (called an alert)

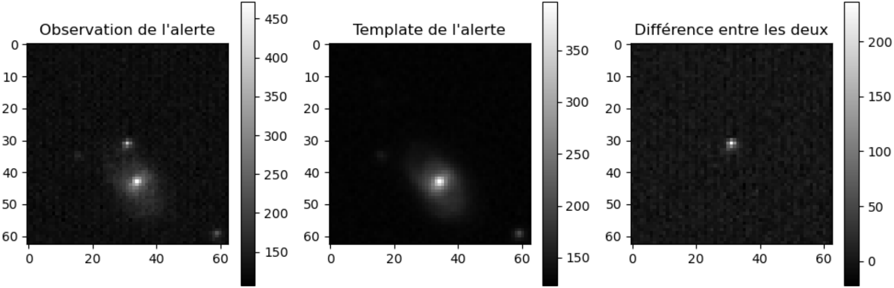
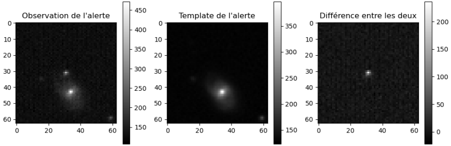


Figure 1: Issuing an alert

The data set

A dataset element (called an alert):

- 110 characteristics specific to alert observation

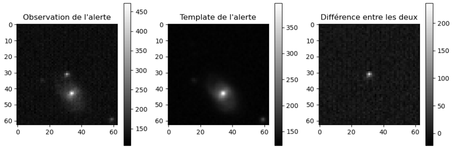


(a) Issuing an alert

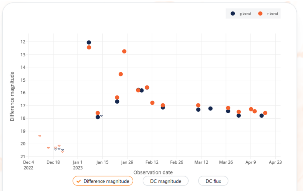
The data set

A dataset element (called an alert):

- 110 characteristics specific to alert observation
- 80 statistical values calculated from the light curve



(a) Issuing an alert

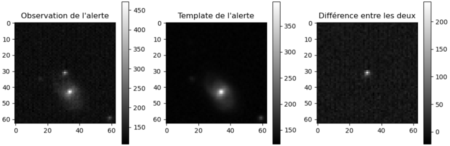


(b) Light curve of an object

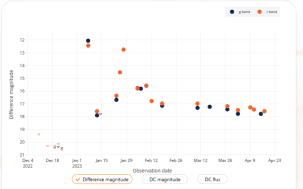
The data set

A dataset element (called an alert):

- 110 characteristics specific to alert observation
- 80 statistical values calculated from the light curve
- 20 values added by Fink including a classification



(a) Issuing an alert

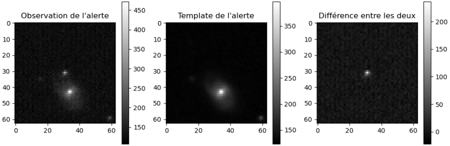


(b) Light curve of an object

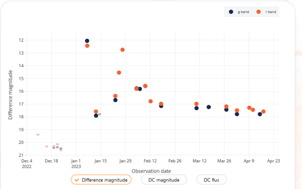
The data set

A dataset element (called an alert):

- 110 characteristics specific to alert observation
- 80 statistical values calculated from the light curve
- 20 values added by Fink including a classification



(a) Issuing an alert



(b) Light curve of an object

Our dataset is composed of 6 distinct classes with 200 elements each

Goal and motivation

- Current situation: traditional tabular representation of alert properties

Goal and motivation

- Current situation: traditional tabular representation of alert properties
- Explore graphs

Goal and motivation

- Current situation: traditional tabular representation of alert properties
- Explore graphs
- Highlight relationships between entities

Goal and motivation

- Current situation: traditional tabular representation of alert properties
- Explore graphs
- Highlight relationships between entities
- Identify anomalies among entities

Dimensionality reduction

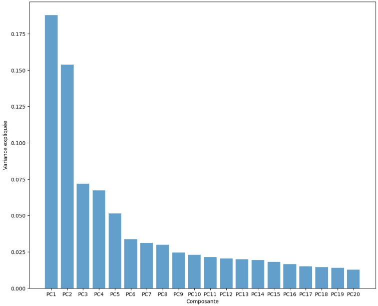
- Hand sorting of features, those relevant to the observation or the position etc ..

Dimensionality reduction

- Hand sorting of features, those relevant to the observation or the position etc ..
- Principal analysis component after converting and normalizing features

Dimensionality reduction

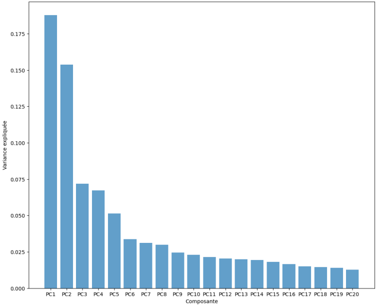
- Hand sorting of features, those relevant to the observation or the position etc ..
- Principal analysis component after converting and normalizing features



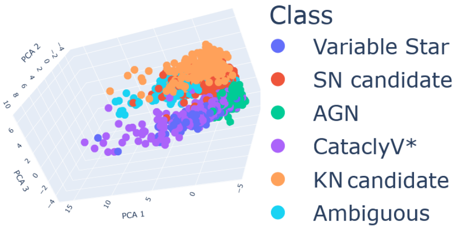
(a) Variance explained by principal component

Dimensionality reduction

- Hand sorting of features, those relevant to the observation or the position etc ..
- Principal analysis component after converting and normalizing features



(a) Variance explained by principal component



(b) Alerts based on the first three components

Graphs

Definition 1

A graph G is defined as an ordered pair of vertex and edges $G = (V, E)$

Graphs

Definition 1

A graph G is defined as an ordered pair of vertex and edges $G = (V, E)$

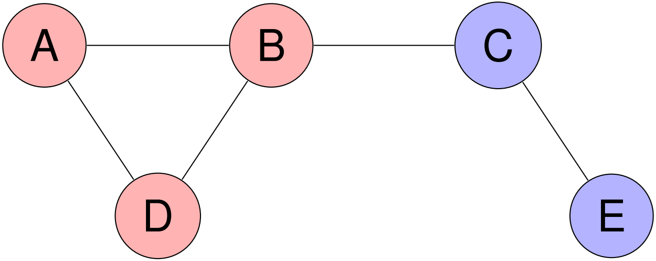


Figure 3: Example of a graph

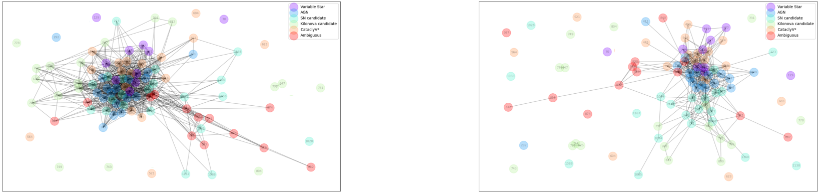
In our case, vertex are alerts but how to define edges ?

Edge definition

For every alert in the sample of our dataset, we create a link for all the alert where the euclidian distance in $\mathbb{R}^{20}$ is inferior to a certain limit r.

Edge definition

For every alert in the sample of our dataset, we create a link for all the alert where the euclidian distance in $\mathbb{R}^{20}$ is inferior to a certain limit r.

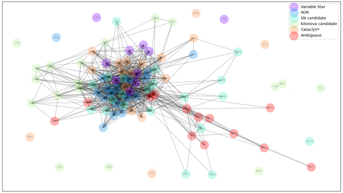


(a) $n\sqrt{n}$

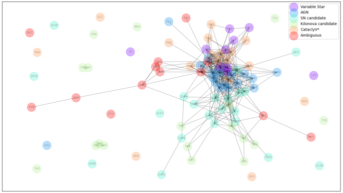
(b) $n \log n$

Edge definition

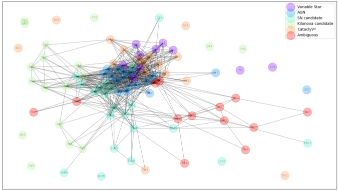
For every alert in the sample of our dataset, we create a link for all the alert where the euclidian distance in $\mathbb{R}^{20}$ is inferior to a certain limit r.



(a) $n\sqrt{n}$



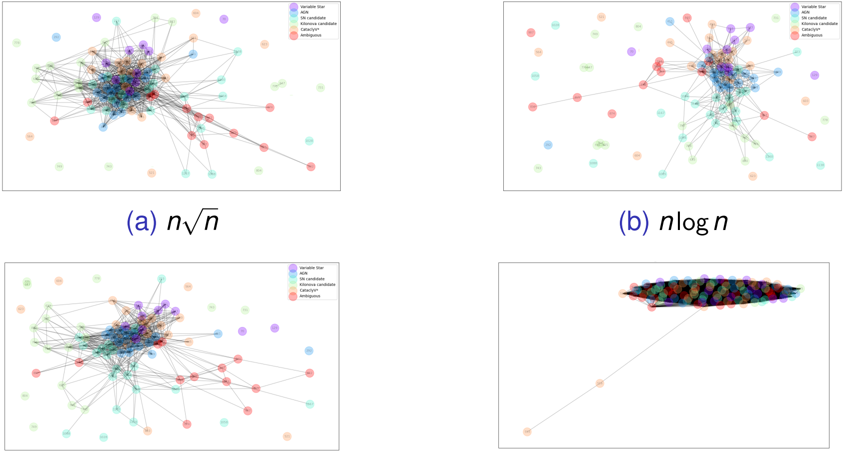
(b) $n\log n$



(c) Theorical number of edges

Edge definition

For every alert in the sample of our dataset, we create a link for all the alert where the euclidian distance in $\mathbb{R}^{20}$ is inferior to a certain limit r.



(a) $n\sqrt{n}$

(b) $n \log n$

(c) Theorical number of edges

(d) One connected component

Performance measurement

Definition 2

A good edge is defined as an edge between two alerts of the same class.

Performance measurement

Definition 2

A good edge is defined as an edge between two alerts of the same class.

Definition 3

The accuracy of a graph G is defined as the ratio of the number of good edges to the total number of edges.

Performance measurement

Definition 2

A good edge is defined as an edge between two alerts of the same class.

Definition 3

The accuracy of a graph G is defined as the ratio of the number of good edges to the total number of edges.

Definition 4

We define the similarity density of a graph G as the ratio of the number of good edges to the total number of hypothetical good edges.

Performance measurement

Definition 2

A good edge is defined as an edge between two alerts of the same class.

Definition 3

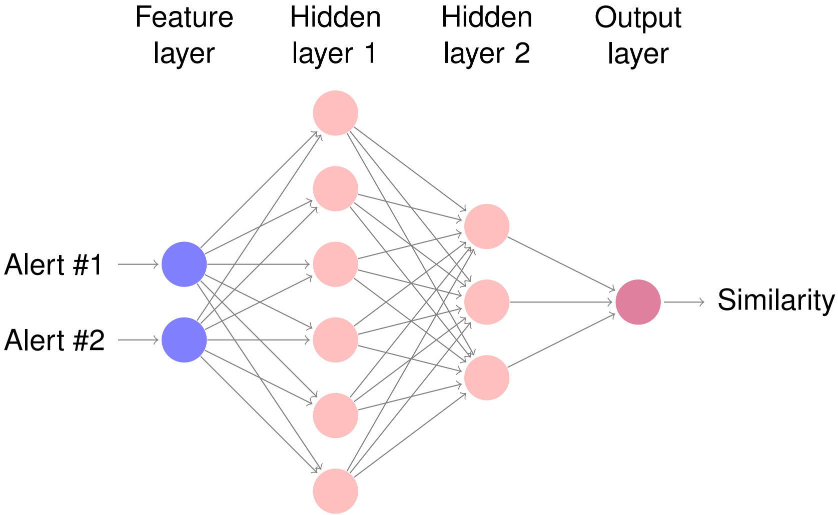
The accuracy of a graph G is defined as the ratio of the number of good edges to the total number of edges.

Definition 4

We define the similarity density of a graph G as the ratio of the number of good edges to the total number of hypothetical good edges.

Type of construction	$n\sqrt{n}$	$n \log n$	TNE	OCC
Accuracy	34%	40%	36%	16%
Similarity density	38%	23%	34%	79%

Neural Network



Technical details: learning rate: 0.001, 200 epochs, Loss function: Binary Cross Entropy, optimizer: Adam
Framework used: Pytorch
Training set: 200 alerts, Test set: 200 alerts

Result

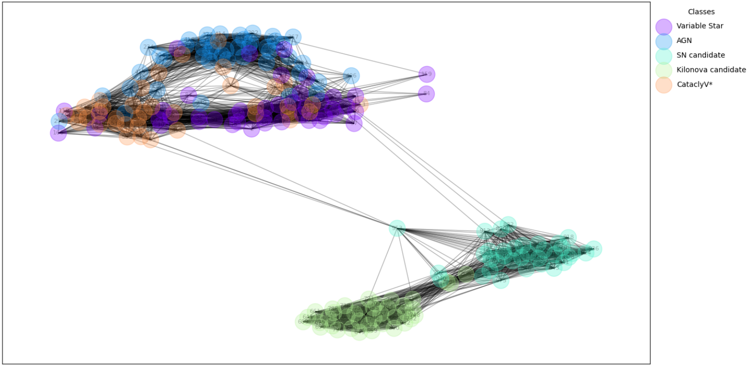


Figure 5: Neural network graph

Result

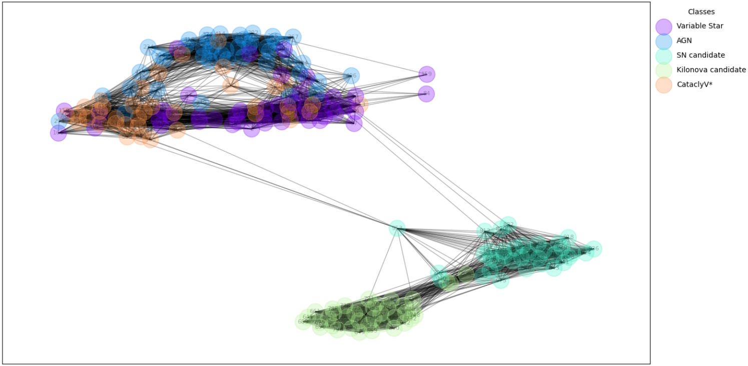


Figure 5: Neural network graph

Accuracy	Similarity density
64%	61%

Result

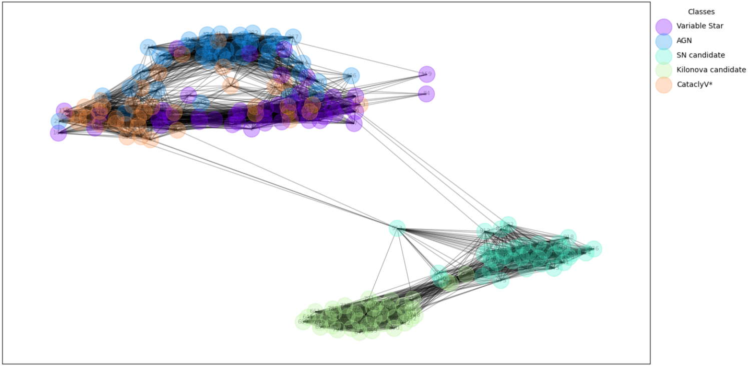


Figure 5: Neural network graph

Accuracy	Similarity density
64%	61%

78% of alerts are linked to a majority of alerts of the same type

What about anomalies ?

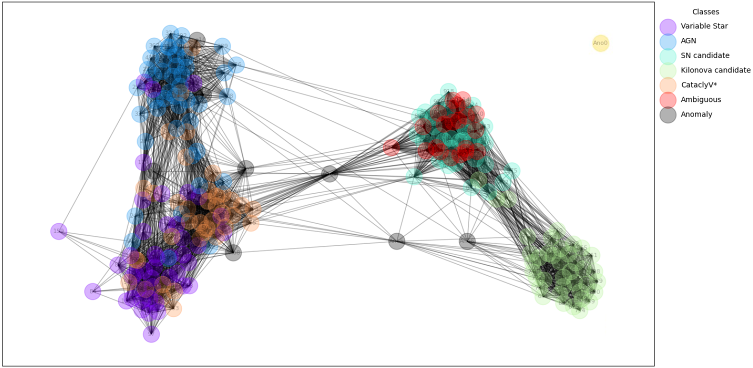


Figure 6: Adding anomalies and ambiguous alerts

What about anomalies ?

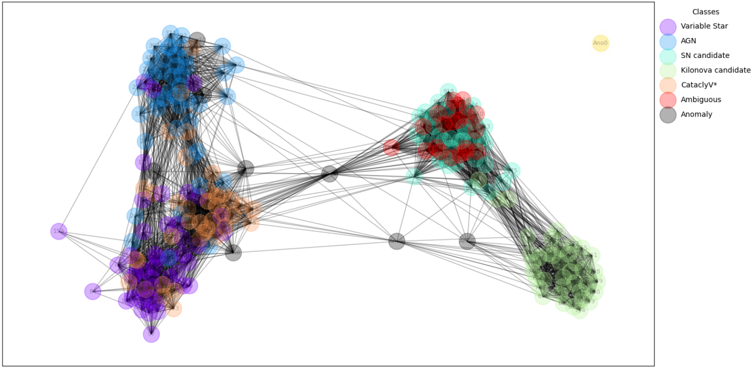


Figure 6: Adding anomalies and ambiguous alerts

- Anomalous alerts are connected in a particular way

What about anomalies ?

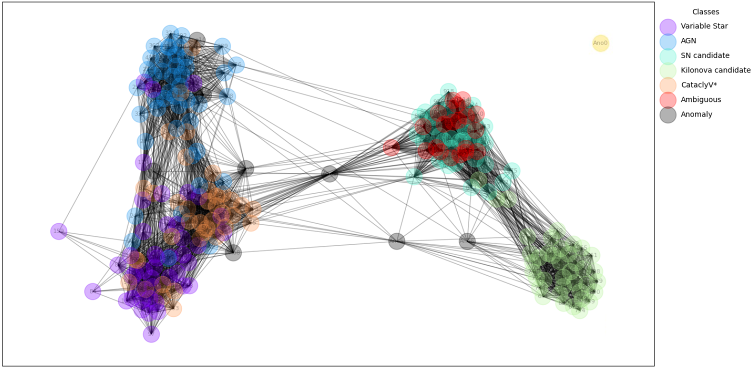


Figure 6: Adding anomalies and ambiguous alerts

- Anomalous alerts are connected in a particular way
- Ambiguous alerts are connected to an average of 90% of supernovas

Perspectives

Potential improvement

- Stabilize results: Impact of the training dataset
- Refine prediction: Graph neural network
- Scaling up: Hypergraphs
- Improving the interpretability of predictions: Modifying the loss function

Perspectives

Potential improvement

- Stabilize results: Impact of the training dataset
- Refine prediction: Graph neural network
- Scaling up: Hypergraphs
- Improving the interpretability of predictions: Modifying the loss function

Short term application

- Recommendation system
- Detection of potential anomalies
- Enhance Fink services for the scientific community

Thank you for your attention
Any questions ?