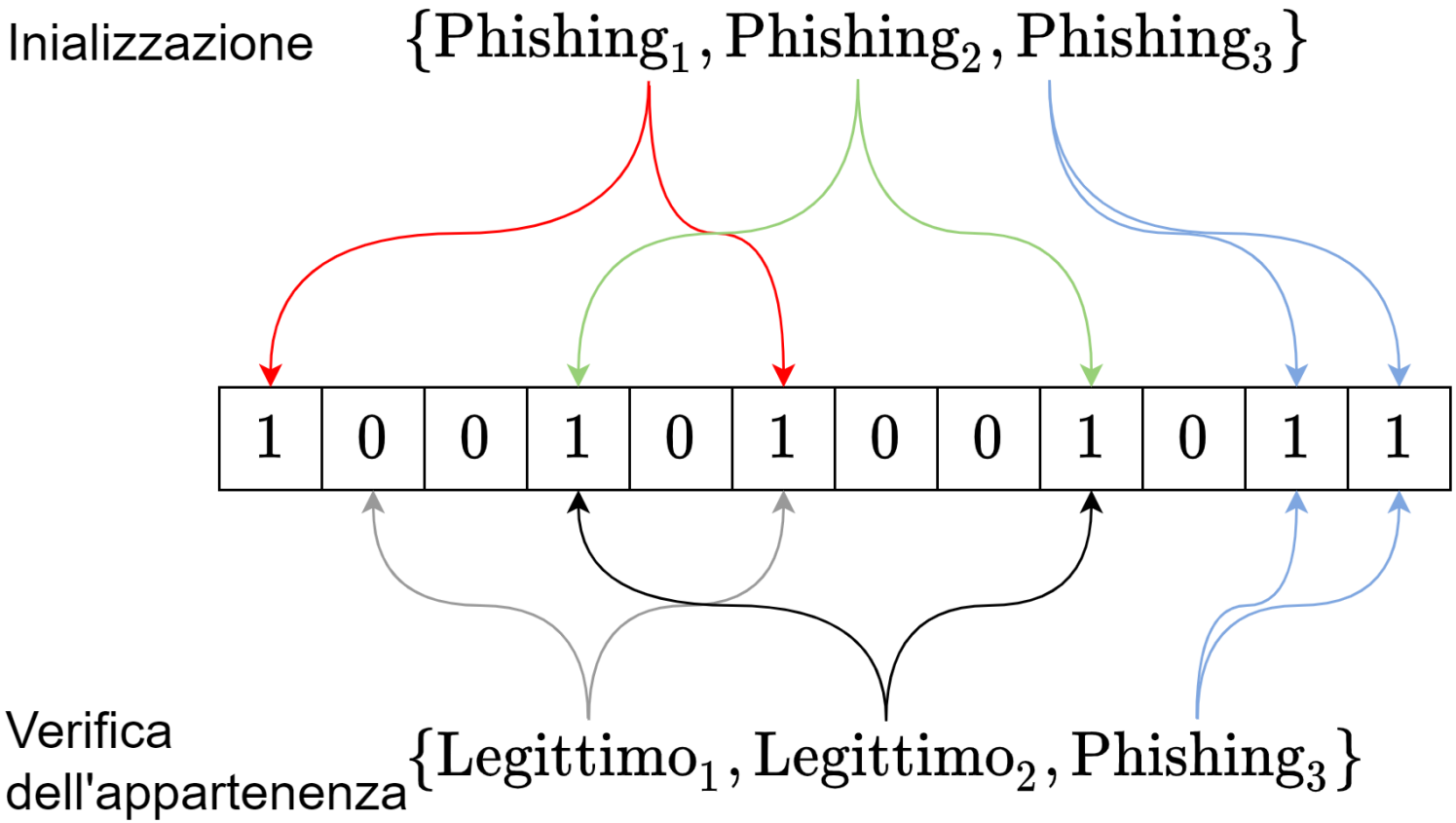
Sulla sinergia tra apprendimento automatico e filtri di Bloom

Giacomo Fumagalli

Relatore: Prof. Dario Malchiodi
Correlatore: Prof. Marco Frasca

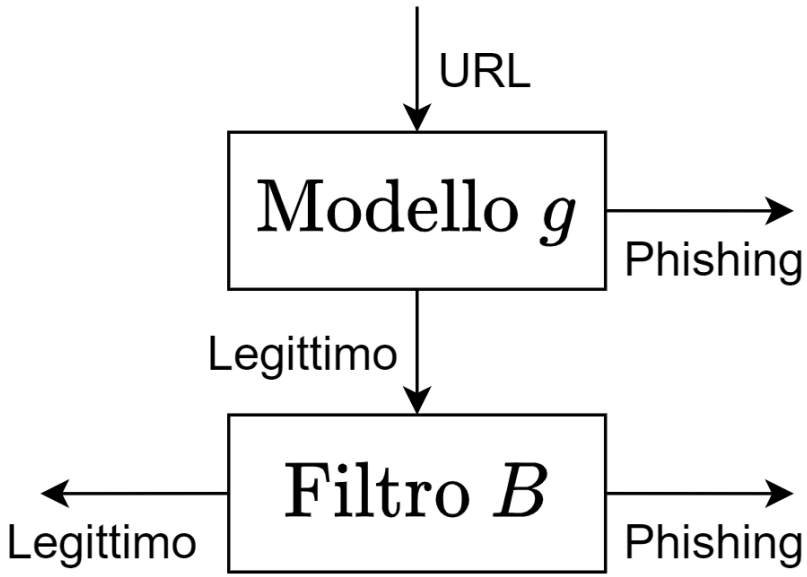
Filtro di Bloom

- Struttura dati basata su funzioni di hash, utile a rappresentare in modo approssimato un insieme di elementi.



Learned Bloom filter (LBF)

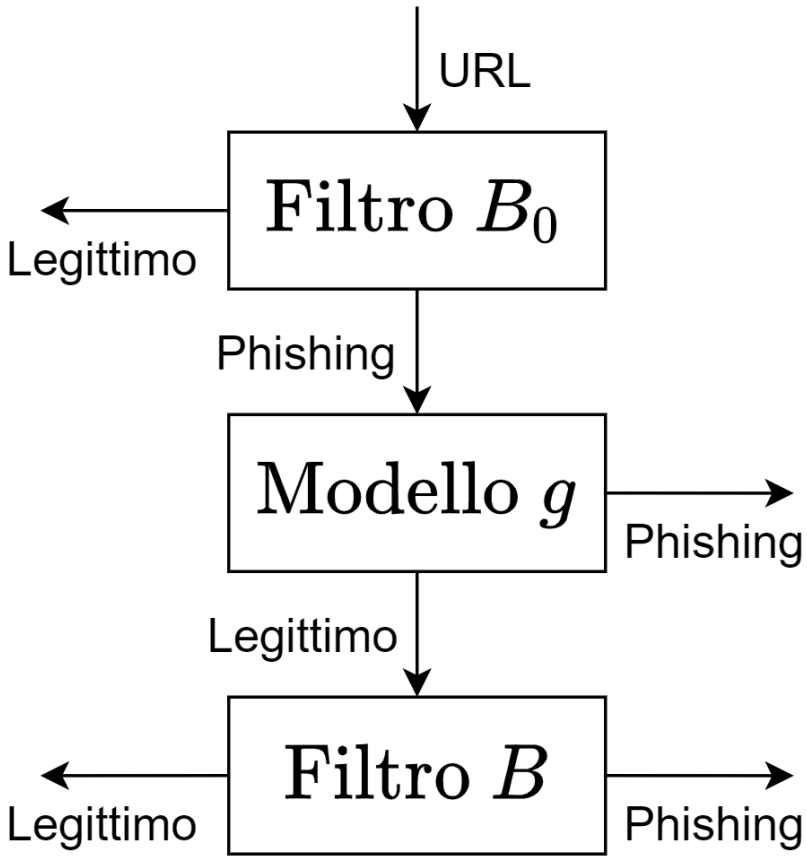
- Composto da: classificatore g, soglia τ, filtro di backup B.
- Riduzione dello spazio grazie a g, che ha la funzione di 'pre-filtro'.



Verifica dell'appartenenza.

Sandwiched learned Bloom filter (SLBF)

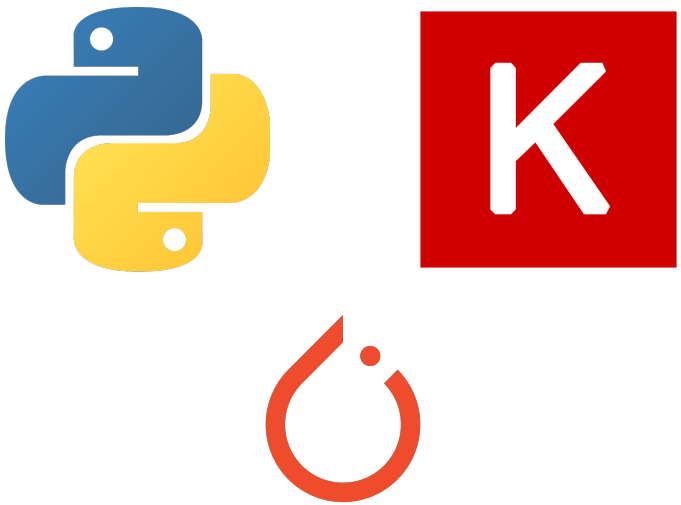
- Composto da: filtro iniziale B_0, classificatore g, soglia τ, filtro di backup B.
- Riduzione dello spazio grazie a B_0, che elimina la maggior parte delle non-chiavi.



Verifica dell'appartenenza.

Esperimenti - Introduzione

- Punto di partenza: "An Empirical Analysis of the Learned Bloom Filter and its Extensions[1]".
- Dataset $\approx$ 310.000 URL:
  - 270.000 URL legittimi,
  - 40.000 URL di phishing.
- Classificatori analizzati: rete ricorrente e percettrone multistrato.



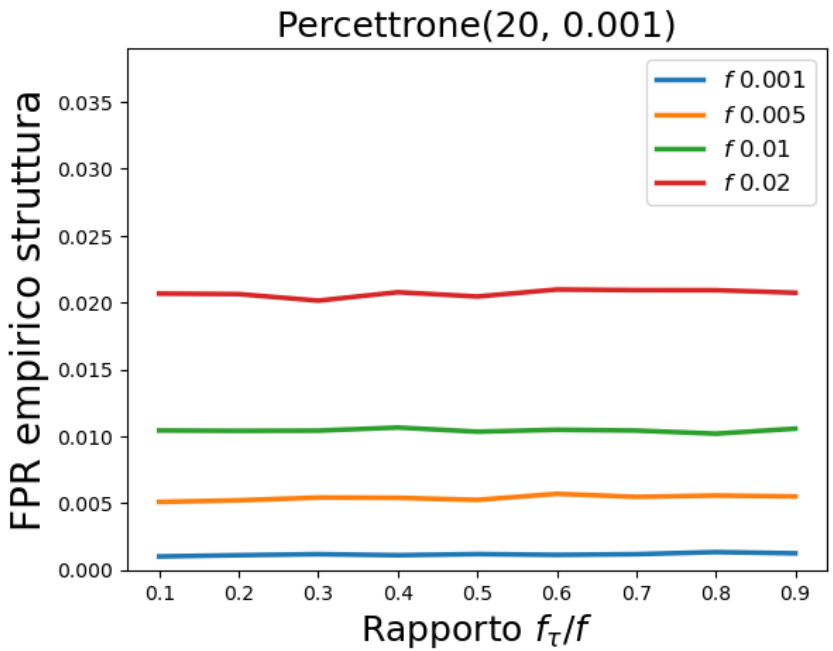
[1] Celine Liang Jason Ma. "An Empirical Analysis of the Learned Bloom Filter and its Extensions". In: (2020).

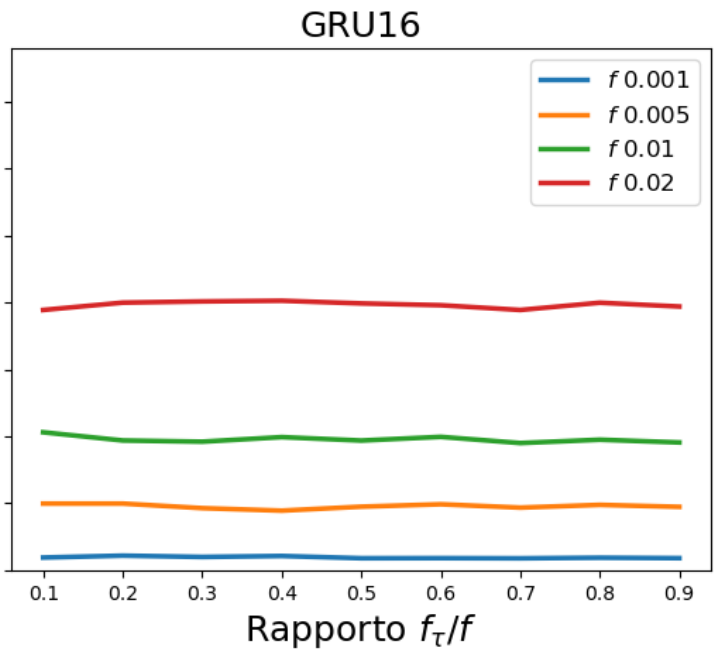
Esperimenti

Due categorie di esperimenti:

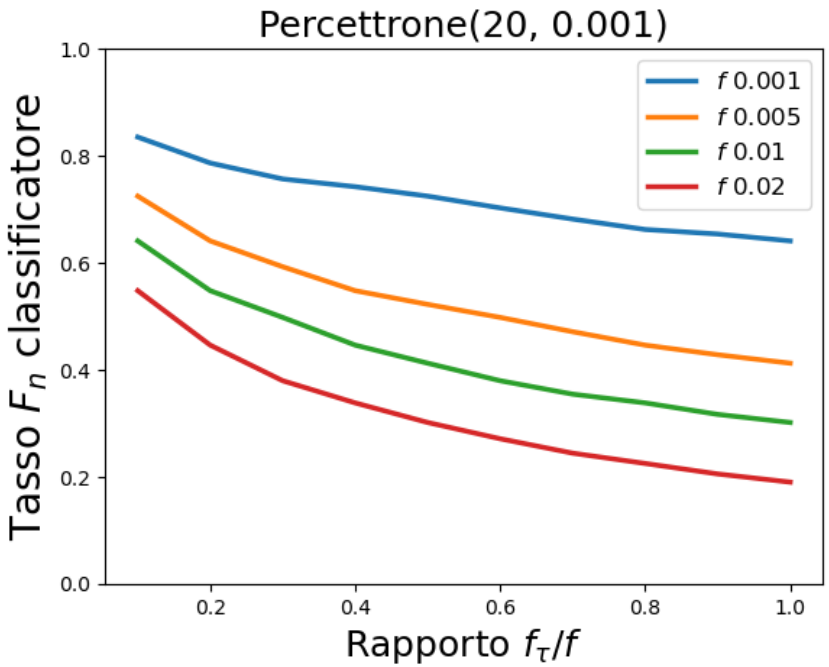
- Esperimenti preliminari:
  - Valutazione delle prestazioni della rete ricorrente.
  - Valutazione delle prestazioni del percettrone multistrato.
- Analisi empiriche di LBF e SLBF su diversi valori di τ:
  - Tasso empirico di falsi positivi della struttura.
  - Tasso di falsi negativi del classificatore.
  - Taglia della struttura.

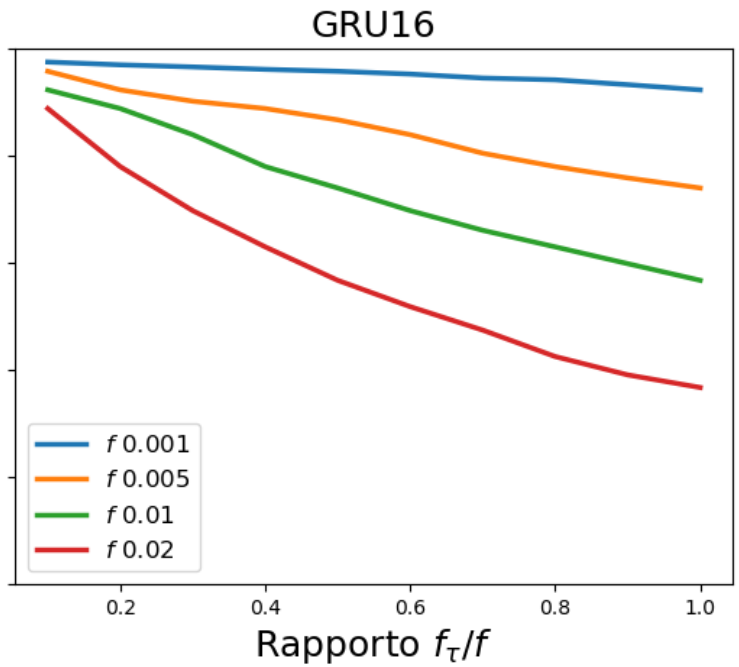
Risultati - Analisi empiriche LBF (1)



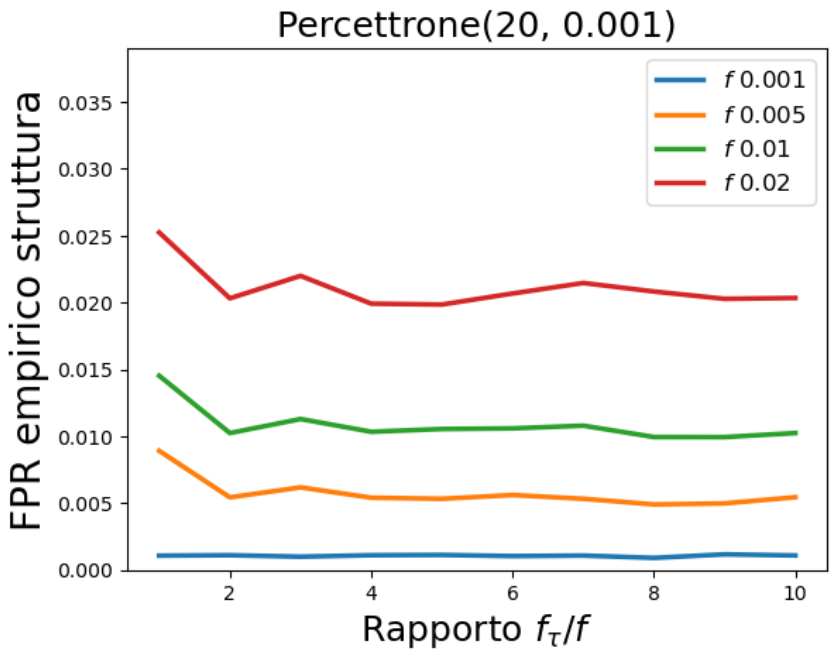


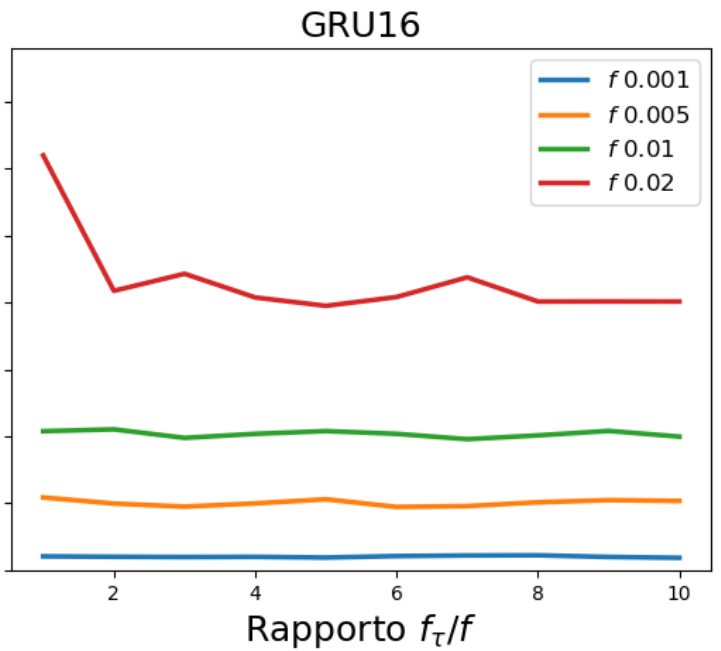
Risultati - Analisi empiriche LBF (2)



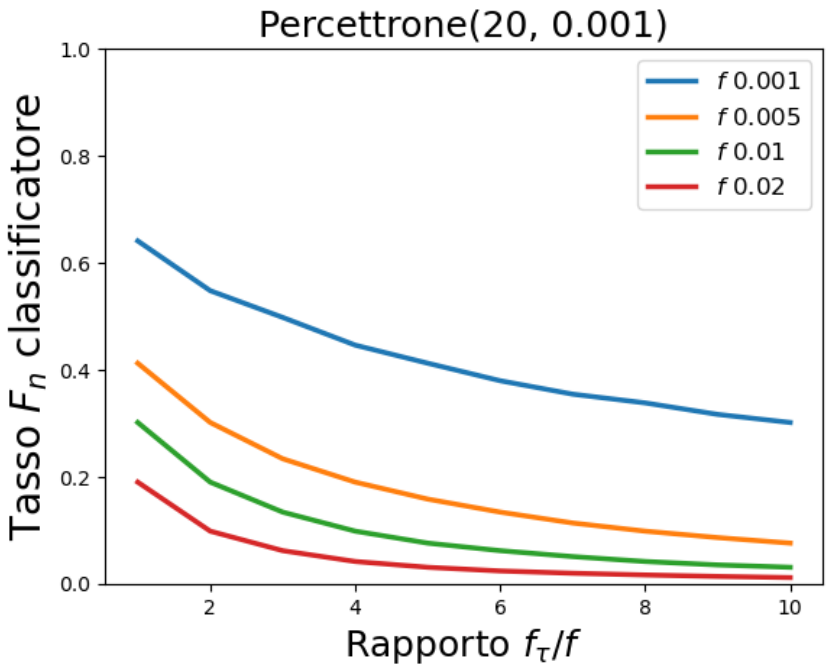


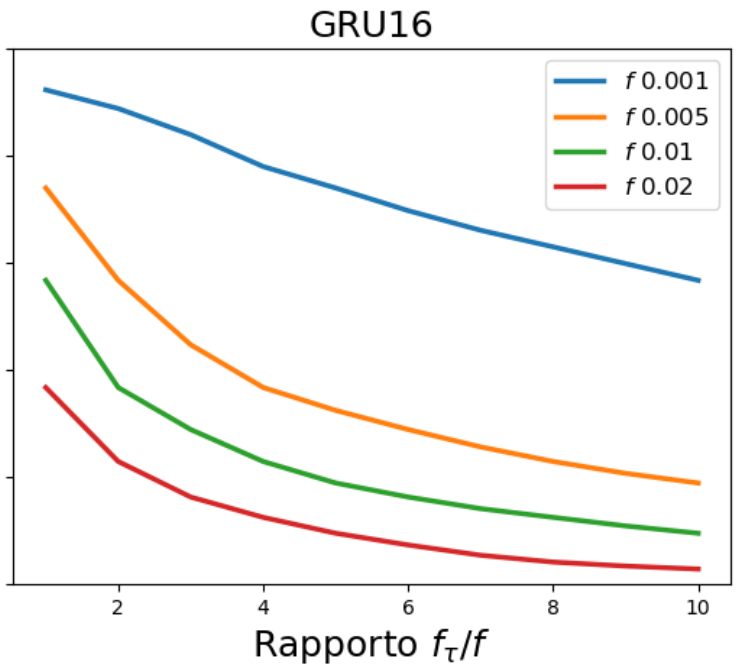
Risultati - Analisi empiriche SLBF (1)



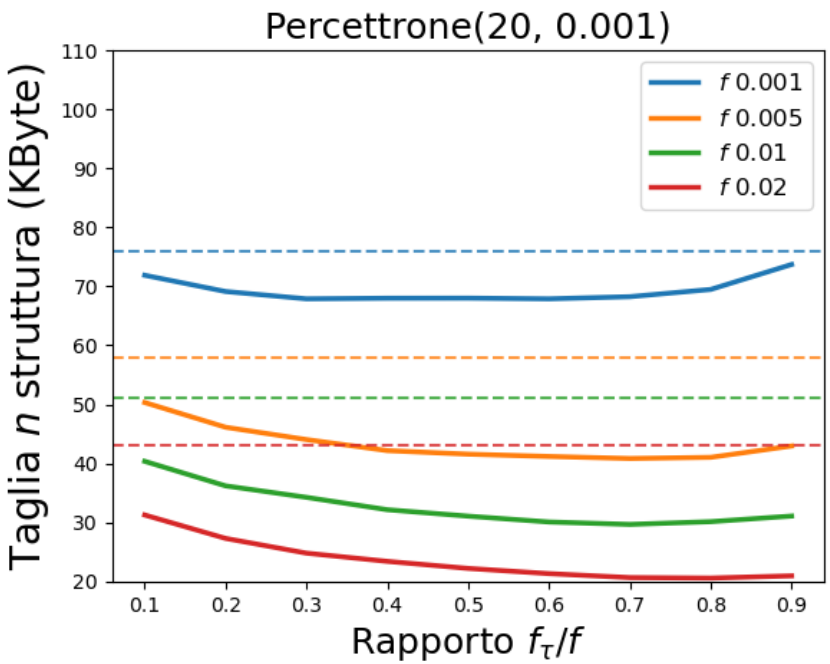


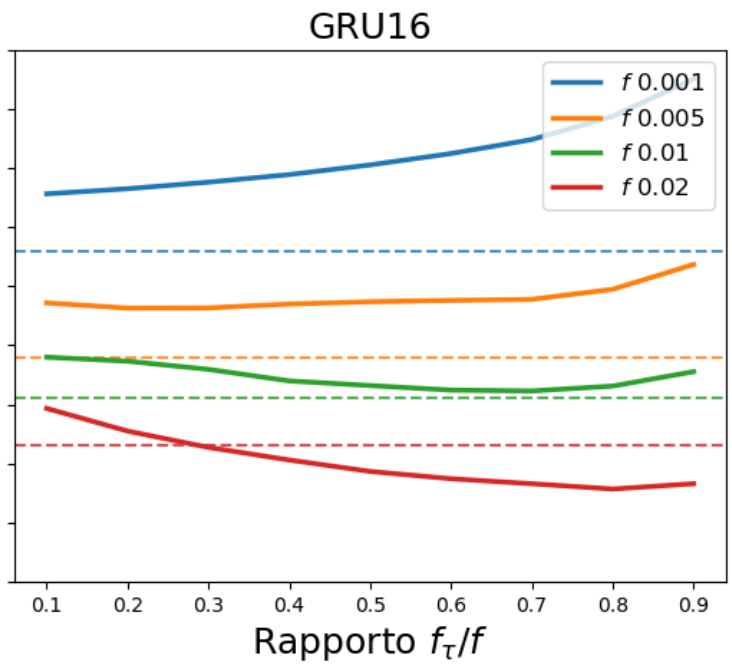
Risultati - Analisi empiriche SLBF (2)



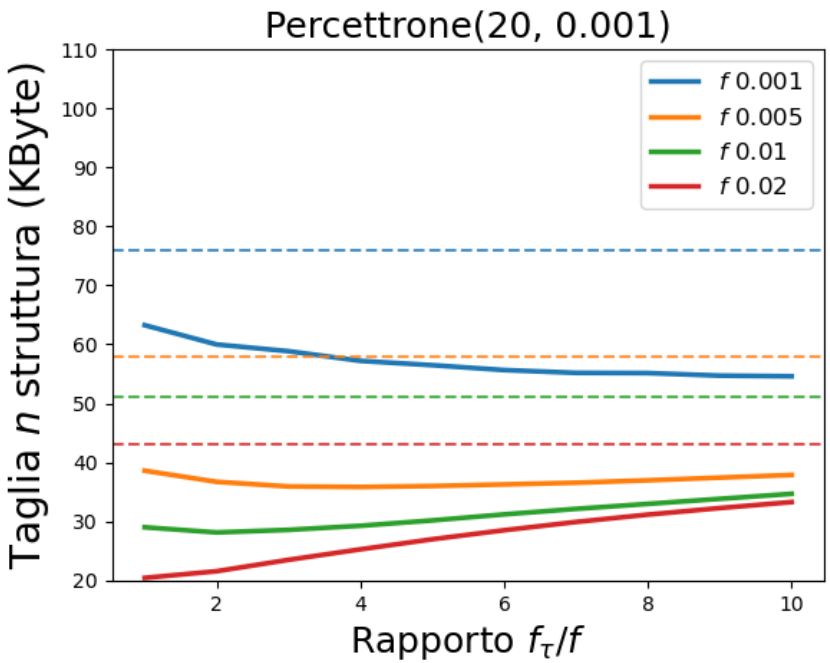


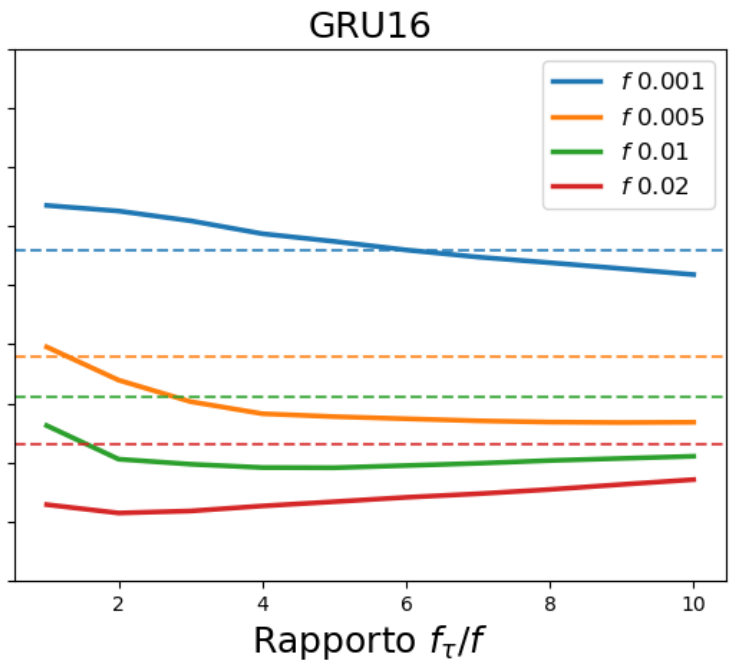
Risultati - Spazio occupato (LBF)





Risultati - Spazio occupato (SLBF)





Conclusioni

- Prestazioni migliori del percettrone nel problema considerato, conseguentemente i filtri di Bloom appresi costruiti con questo modello risultano migliori.
- Possibili sviluppi futuri:
  - Ricerca di criteri ottimali per la scelta di τ.
  - Analisi utilizzando tipologie di classificatore differenti.
  - Analisi utilizzando dataset più grandi.