Analisi dei risultati ottenuti

Costruzione dei dataset e descrizione

Lo scopo di questo esperimento è stato quello di capire cosa può accadere in una situazione realistica in cui uno ha in fase di training molte istanze *normali* e poche di attacco (*anomaly*). Ci ritorneremo sopra, me se da un lato ha sembrato avere risultati migliori, dall'altro no.

1. Questa volta per prima cosa sono state riunite in un unico file tutte le classi *normal* e solo il 20% di classi *anomaly* KDD_AllNormal_20%Anomaly.arff, questo file a sua volta è stato filtrato tramite la feature selection che ha prodotto il file:

KDD_AllNormal_20%Anomaly_Filtered.arff

- 2. Succesivamente si è proceduto nella maniera analoga a prima, si è estratto l'80% delle classi normal, e si è creato il file KDD_Train_80%Normal_Filtered da usare come train set.
- 3. Per creare il test set, si è preso il restante 20% delle classi normal, e si è unito con con tutte le classi anomaly presenti sia nel file di Train e Test ufficiali dati da NSL (KDDTrain+.arff e KDDTest+.arff) avendo l'accortezza di eliminare gli attributi che erano stati scartati con la selezione, il file risultante si chiama:

KDD_Test_20%NormalAllAnomaly

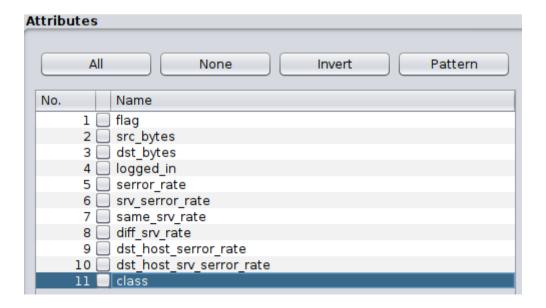
Risultati ottenuti

Il file KDD_AllNormal_20%Anomaly_Filtered.arff contiene:

- 77054 istanze normali
- 14293 istanze anomaly

di conseguenza la selezione degli attributi ha preso in considerazione questo numero di istanze per le due classi, ed il risultato degli attributi selezioni è stato differente:

Attributi selezionati



Train_set	Test_set
KDD_Train_80%Normal_Filtered	KDD_Test_20%NormalAllAnomaly

I risultati ottenuti sono presenti nel file Analisi.csv, qui di seguito vi è riportato uno screenshot della tabella:

File	%Correttezza	%Anomaly	%FN	Istanze_?_nel_test	Istanze_normal_nel_tes	Tot_istanze_nel_test	%FP
Train_NormalFiltered_Test_20%NormalAllAnomalyFiltered_csv	83.63	96.57	3.43	71463	15411	86874	76.36

Considerazioni

Da una parte l'esperimento ha prodotto risultati migliori, ad esempio, nonostante il numero di istanze *anomaly* sia stato inferiore, la percentuale di riconoscimento delle anomalie è aumento di quasi il 10% avendo come conseguenza anche un calo dei falsi negativi.

Tuttavia, la percentuale di falsi positivi è aumentata drasticamente, di consguenza quasi 8 situazioni normali su 10 vengono etichettate come una anomalia.