Analisi dei risultati ottenuti

Costruzione dei dataset e descrizione

Partendo dal file Train e Test originali, ho proceduto nel seguente modo:

1. Ho creato un file per l'addestramento contenente unicamente le istanze *normal* contenute sia nel file `KDDTrain+.arff` che `KDDTest+.arff`, questo file è stato chiamato:

   KDDTrain_OnlyNormal

2. Ho creato un dataset per il testing contenente unicamente le istanze *anormali* contenute sia nel file `KDDTrain+.arff` che `KDDTest+.arff`, questo file è stato chiamato:

   KDDTest_OnlyAnomaly

3. Successivamente ho realizzato una versione ridotta dei due file, chiamati rispettivamente:

   KDDTrain_OnlyNormal20% e KDDTest_OnlyAnomaly20%

   contenenti il 20% delle istanze totali

4. Infine, a partire da questo ultimo file di test, avente il 20% delle istanze totali, ho deciso di suddividerle in un ulteriore 80% di classe *anomaly* e 20% *normal*, il file prende il nome di:

   KDDTest_20%Normal80%Anomaly

Risultati ottenuti

Ho deciso di testare i seguenti abbinamenti:

Train_set	**Test_set**
KDDTrain_OnlyNormal	KDDTest_OnlyAnomaly
KDDTrain_OnlyNormal20%	KDDTest_OnlyAnomaly20%
KDDTrain_OnlyNormal20%	KDDTest_20%Normal80%Anomaly
KDDTrain_OnlyNormal	KDDTest_20%Normal80%Anomaly

I risultati ottenuti sono presenti nel file `Analisi_csv`, qui di seguito è riportato uno screenshot della tabella:

File	%Correttezza	%Anomaly	%FN	Istanze ? nel test	Istanze normal nel test	Tot istanze nel test	%FP
FullTrain_FullTest_csv	-	98.66	1.34	71463	0	71463	-
TrainNormal20_TestAnomaly20_csv	-	98.94	1.06	14294	0	14294	-
TrainNormal20_TestNormal20Anomaly80_csv	89	98.93	1.07	11389	2860	14249	50
FullTrainNormal_TestNormal20Anomaly80_csv	88.85	98.98	1.02	11389	2860	14249	51

Considerazioni

Il rilevamento delle anomalie sembra funzionare molto bene, dato che in tutti i test presenta una percentuale di accuratezza maggiore del 98%.

Tuttavia, la percentuale dei falsi positivi (FP, cioè istanze normali che vengono classificate come anomalie) è molto alta (del 50%).
L'ultimo test aveva come obbiettivo quello di addestrare meglio l'algoritmo con la speranza di diminuire questa percentuale, ma non è stato ottenuto il risultato sperato.