The initial experiment is performed using K-means clustering in spark.

The Data used to formed clusters:

Number of Training Data Samples: 4898431

Number of Testing data Samples: 311029

For testing data the attack types count:

smurf. 164091

normal. 60593

neptune. 58001

snmpgetattack. 7741

mailbomb. 5000

guess_passwd. 4367

snmpguess. 2406

satan. 1633

warezmaster. 1602

back. 1098

mscan. 1053

apache2. 794

processtable. 759

saint. 736

portsweep. 354

ipsweep. 306

httptunnel. 158

pod. 87

nmap. 84

buffer_overflow. 22

multihop. 18

sendmail. 17

named. 17

ps. 16

rootkit. 13

xterm. 13

teardrop. 12

xlock. 9

land. 9

xsnoop. 4

ftp_write. 3

phf. 2

worm. 2

sqlattack. 2

loadmodule. 2

perl. 2

udpstorm. 2

imap. 1

Clustering in different setup: Also 3 features (Symbolic Features are dropped) Why?  Read in one of the website dealing with kddcup1999 data for clustering they are also removing the symbolic features.

With number of clusters: 23

cluster 0 : 152317 +   154371

cluster 1 :

cluster 2 :

cluster 3 :

cluster 4 :

cluster 5 :

cluster 6 :

cluster 7 :

cluster 8 : 1

cluster 9 : 2

cluster 10:

cluster 11:

cluster 12: 2

cluster 13:

cluster 14: 12

cluster 15: 326 +  197

cluster 16: 1

cluster 17: 1

cluster 18: 1

cluster 19: 1029

cluster 20: 10 + 28

cluster 21:

cluster 22:

With number of clusters as 23 the results does not look convincing. So I wrote a logic to get the cluster score in different value of K (Number of clusters) and k varies from 10 to 23. After doing this the value suggested is : 20  . Then clustering is performed with value as 20.

cluster 0 : 155367 + 155598

cluster 1 :

cluster 2 :

cluster 3 :

cluster 4 :

cluster 5 :

cluster 6 :

cluster 7 :

cluster 8 : 1

cluster 9 : 1

cluster 10:

cluster 11:

cluster 12: 3

cluster 13:

cluster 14: 2

cluster 15: 43

cluster 16:

cluster 17: 11

cluster 18: 2

cluster 19:

Still the results are not good. Looking into the reason.