训练集：kdd10%的训练集（kddcup训练集）为训练集1

测试集：corrected10%的数据集

数据集中的数据类别：

1）正常状态(Normal)：Normal

2）拒绝服务攻击(Dos,denial-of-service)：Back, Land, Neptun , Pod, Smurf, Teardrop, Mailbomb, Processtable, Udpstorm, Apache2, Worm

3）远程攻击(R2L,unauthorized access from a remote machine)：Guess\_passwd, Ftp\_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Xlock,Xsnoop, Snmpguess, Snmpgetattack, Httptunnel

4）提权攻击(U2R,unauthorized access to local superuser (root) privileges)：Satan,I Psweep, Nmap, Portsweep, Mscan, Saint

5）端口扫描(Probing,surveillance and other probing)：Buffer\_overflow, Loadmodul, Rootkit, Perl, Sqlattack, Xterm, Ps

1. 程序1

用*K-means*改进的SMOTE技术对U2R进行合成少数类别样本，将U2R的样本数增加到500条，添加到训练集1，得到新的训练集2.

即用*K-means*改进SMOTE技术，对U2R进行合成少数类别样本，即用*K-means*计算出样本点的簇心，选取簇心的*k*个近邻进行插值操作，得出新样本。

1. 程序2 (该部分可参考文件夹中，程序replace\_string\_to\_value.py和processing.py是可执行程序)

数据归一化

字符型数据数字化

数字型数据归一化

对两个训练集1、2和测试集标签类别分两类：

1. 二个标签

Normal 0；其他 1

1. 多标签

Normal 1；Dos 2；R2L 3；U2R 4；Probing 5

形成四个训练集未经改进SMOTE处理的二分类和多分类数据集；经过改进SMOTE处理的二分类和多分类数据集。

两个测试集二分类测试集和多分类测试集。

1. 程序3

将训练集输入到深度学习网络中（要求可以修改数据的维度），训练DBN网络，BP微调模型参数，将高维数据映射为低维数据集。用softmax逻辑回归进行二分类或多分类，输入训练集验证结果。

1. 程序4

将训练集输入到深度学习网络中，训练DBN网络，BP微调模型参数，进行二分类或多分类，输入训练集验证结果。

1. 实验输出

（1）经过SMOTE处理的数据检测率较高：

二分类：经过改进SMOTE技术处理前后的训练集，Normal被划分为0的个数，被划分为1的个数

其他异常类被划分为0的个数，被划分为1的个数

多分类：经过改进SMOTE技术处理前后的训练集，Normal分别被划分为1-5类的个数；Dos分别被划分为1-5类的个数；R2L 分别被划分为1-5类的个数；U2R 分别被划分为1-5类的个数；Probing分别被划分为1-5类的个数。展示结果如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Normal | Dos | R2L | U2R | Probing |
| Normal |  |  |  |  |  |
| Dos |  |  |  |  |  |
| R2L |  |  |  |  |  |
| U2R |  |  |  |  |  |
| Probing |  |  |  |  |  |

1. 检测率和误报率

预期结果：

1）经过SMOTE处理的数据集训练网络比未经过处理的数据集训练网络，得到的检测率高

2）经过SMOTE处理的数据集训练网络比未经过处理的数据集训练网络，少数类别样本（U2R）的检测率提高