Experiment 1

Dataset Preparation and Analysis

Examined both datasets to evaluate the percentage of normal records in the dataset. The objective was to ensure that the dataset was highly imbalanced to reflect an operational network that was not compromised simultaneously with large numbers of attacks. In the initial analysis, the NSL-KDD dataset was found to have only 53% normal records. This was largely because of the many denial of service attacks contained in the dataset and because many of the duplicate normal records were removed in the creation of this dataset over the original KDD Cup 1999 dataset. Table 1 shows the percentage of each class of records in the NSL-KDD dataset:

|  |  |  |
| --- | --- | --- |
| Label | Records | PctTotal |
| normal | 67,343 | 53% |
| neptune | 41,214 | 33% |
| satan | 3,633 | 3% |
| ipsweep | 3,599 | 3% |
| portsweep | 2,931 | 2% |
| smurf | 2,646 | 2% |
| nmap | 1,493 | 1% |
| back | 956 | 1% |
| teardrop | 892 | 1% |
| warezclient | 890 | 1% |
| pod | 201 | 0% |
| guess\_passwd | 53 | 0% |
| buffer\_overflow | 30 | 0% |
| warezmaster | 20 | 0% |
| land | 18 | 0% |
| imap | 11 | 0% |
| rootkit | 10 | 0% |
| loadmodule | 9 | 0% |
| ftp\_write | 8 | 0% |
| multihop | 7 | 0% |
| phf | 4 | 0% |
| perl | 3 | 0% |
| spy | 2 | 0% |
| Total | 125,973 |  |

To prepare this dataset to reflect a more realistic operational network, all of the denial of service attacks were removed. The resulting dataset contained 98% normal records. Table 2 shows the classes that remained in the new dataset.

|  |  |  |
| --- | --- | --- |
| Label | Records | PctTotal |
| normal | 67,343 | 98% |
| warezclient | 890 | 1% |
| guess\_passwd | 53 | 0% |
| buffer\_overflow | 30 | 0% |
| warezmaster | 20 | 0% |
| land | 18 | 0% |
| imap | 11 | 0% |
| rootkit | 10 | 0% |
| loadmodule | 9 | 0% |
| ftp\_write | 8 | 0% |
| multihop | 7 | 0% |
| phf | 4 | 0% |
| perl | 3 | 0% |
| spy | 2 | 0% |
| Total | 68,408 |  |

The UNSW-NB15 dataset contains four files of different datasets. Each file contains a different ratio of normal and attack records. Table 3 shows the distribution of normal records as a percentage of abnormal records for these datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| File | Normal Records | Total Records | Pct Total |
| UNSW-NB15\_1 | 677,786 | 700,001 | 97% |
| UNSW-NB15\_2 | 647,252 | 700,001 | 92% |
| UNSW-NB15\_3 | 542,576 | 700,001 | 78% |
| UNSW-NB15\_4 | 351,150 | 440,044 | 80% |

For this research, UNSW-NB15\_1 will be used without modification, since it demonstrates the most imbalanced dataset without modification. Table 4 shows how the normal and attack records are distributed.

|  |  |  |
| --- | --- | --- |
|  | UNSW-NB15\_1 | |
| Label | Records | PctTotal |
| Normal | 677,786 | 97% |
| Generic | 7,522 | 1% |
| Exploits | 5,409 | 1% |
| Fuzzers | 5,051 | 1% |
| Reconnaissance | 1,759 | 0% |
| DoS | 1,167 | 0% |
| Backdoors | 534 | 0% |
| Analysis | 526 | 0% |
| Shellcode | 223 | 0% |
| Worms | 24 | 0% |
| Total | 700,001 |  |

The UNSW-NB15 dataset includes 45 computer systems that are identified with either a source or destination IP address. Attack records are from four computer systems (175.45.176.0, 175.45.176.1, 175.45.176.2, and 175.45.176.3). These four computer systems are used to attack approximately 10 computer systems (149.171.126.10, 149.171.126.11, 149.171.126.12, 149.171.126.13, 149.171.126.14, 149.171.126.15, 149.171.126.16, 149.171.126.17, 149.171.126.18, 149.171.126.19) with a variety of attack types, depending on which file is used.

Coding/Algorithm

The construction of this algorithm used two classes that were developed for this research. The package dataset.py was used to load the dataset from comma separated value (CSV) files, perform preprocessing, and return a Pandas DataFrame that represents the preprocessed data. In addition, the dataset.py file returns the original, unprocessed data in a DataFrame to allow comparison of the results of the clustering ensembles to the original labels. The second package used in this research was clusterer.py. This package was responsible for generating the partitions, given a number of partitions to generate, the minimum feature ratio, the maximum feature ratio, the minimum number of clusters, and the maximum number of clusters.

Parameter Selection

For this experiment, the minimum feature ratio was set to 0.25, which meant that the minimum number of features that were used in the bagging was 0.25 times the number of features in the dataset. The maximum feature ratio was set to 0.75, which meant that the maximum number of features that were used in the bagging was 0.75 times the number of features in the dataset. In addition, the minimum number of clusters was set to three, and the maximum to 20. All of these numbers were arbitrarily selected for this experiment.

Throughout the testing, the range of the number of clusters generated in each partition was evaluated and updated. Each partition that was generated included a pseudorandom number of clusters with a minimum and a maximum value. Initial evaluation runs included a minimum of three and a maximum of 20 clusters. Sampling and visual analysis of the clustering results showed high concentrations of normal records in all of the clusters in the sampled partitions. As a result, the maximum number of clusters partition was updated to 40, but the results were similar. Finally, updating the minimum to 40 and the maximum to 100 showed a better distribution of results. There were still normal records in most clusters, but with this range, the clustering results demonstrated improved discrimination between types of normal records.

Testing

An initial set of 100 partitions was created using the NSL-KDD dataset. This experiment first used visual inspection of the clustering ensemble results by creating a pivot table in Excel for randomly sampled partitions. During this testing, labeled data was used to evaluate the clustering results. Analysis found that there were very few cases in which anomalies clustered cleanly together without normal records. Thus, the initial assumption that anomalies could be found as K standard deviations below the mean number of results in each cluster was found to be unsuccessful. Additionally, the results that were K standard deviations above the mean were found not to be anomalous, but instead, with a high degree of accuracy, were found to be the normal records.

Since this preliminary conclusion was the result of sampling, it was necessary to review all of the partitions to determine if this observation was consistent. To verify this, an algorithm was developed to test each partition, with a goal of determining if the clustering results that were K standard deviations above the mean number of results in each cluster was consistently accurate. The following algorithm was developed to perform this testing:

Load the dataset

Preprocess the dataset

Generate partitions using k-means clustering

Append labels from partitions to original dataset as columns

For each partition:

Create a pivot table with label as column headings and count of records as values

Select rows where the count of records in the cluster were K standard deviations higher than other clusters in the same partition

Save resulting rows in CSV file

The results of the unsupervised algorithm on the NSL-KDD dataset ranged from identifying from 96.1 to 100 percent of normal records in the cluster that had a count of records greater than two standard deviations above the mean. The average was 99.9 percent, which was sufficiently higher than the ratio of normal records to the total records in the dataset, which was 98 percent.

To determine if this algorithm was generalizable beyond the NSL-KDD dataset, the same algorithm was used on the UNSW-NB15\_1 dataset. In this case, the unsupervised algorithm identified from 91.6 to 100 percent of the records, with an average of 98.8 percent. Similarly, this was higher than the 97 percent in the overall dataset.

These results were tested using a variety of numbers of clusters as well as a variety of values for the number of standard deviations above the mean. The most accurate results were found using a range of 40 to 100 clusters. To isolate the normal records, testing found that four standard deviations above the mean number of clusters was the most accurate.

The result of this experiment shows that this approach to unsupervised identification of normal records in highly imbalanced datasets may be useful identifying anomalous records by the process of elimination. As a result, these values will be carried forward in to the subsequent experiments.