Shukai Liu
Email: hj19989@bristol.ac.uk
Course: EMATM0044: Introduction to Artificial Intelligence

Coursework 1
ID: 1909499
Term: TB2 2019

Problem 1
Determine how many distinct source IP addresses, destination IP addresses, and classifications there are in this dataset.

Solution:

- The number of distinct source IP addresses is 98
- The number of distinct destination IP addresses is 261
- The number of distinct classifications is 3

Problem 2
Write code to count the number of records containing each source IP address and each destination IP address. Generate histograms to visualise your results.

Solution:
The number of appearances of each source IP and destination IP is shown in Figure 1 and Figure 2 respectively.

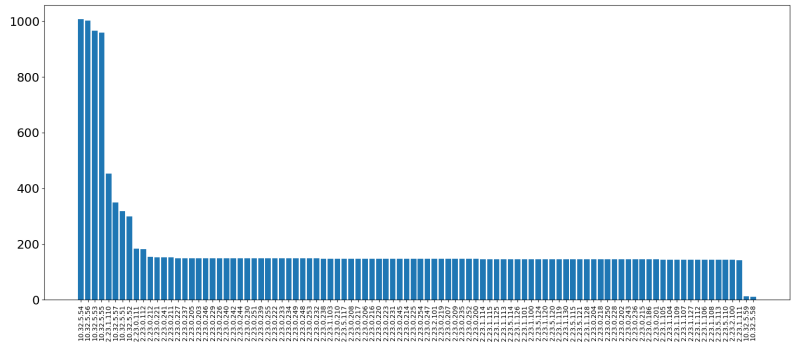


Figure 1: The number of occurrences of source IP

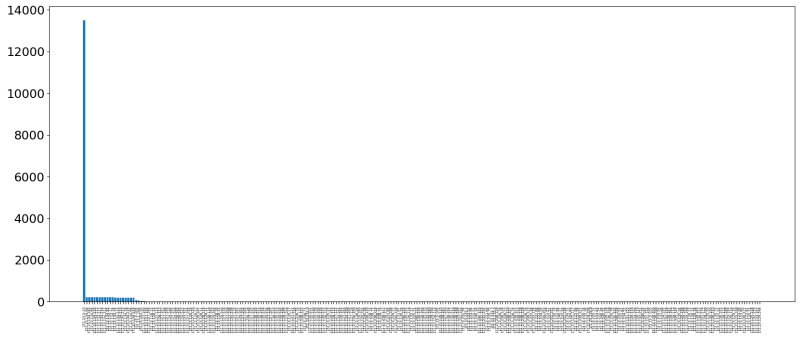


Figure 2: The number of occurrences of destination IP

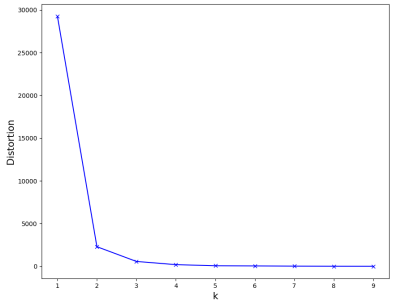
The most frequently occurring source IP address is 10.32.5.54, it appears 1008 times. The most frequent destination IP address is 172.23.0.10 (13505 times), except it, the number of occurrences of other destination IP addresses is less than 250.

Problem 3
ClusteringUsing different method to cluster the source IP addresses and destination IP addresses by the number of records they appear in.

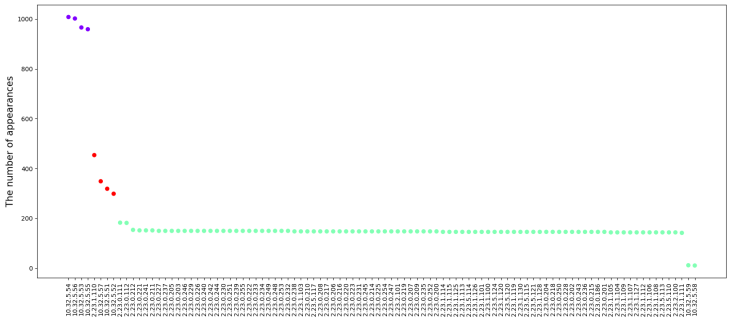
Solution: The data to be classified is the distinct addresses of source IP and destination IP, each of them only has one feature, which means that the data types we need to classify in the program are two one-dimensional lists. Three methods will be used to cluster the data as follow. Because each data has only one feature, so in the graph showing the results, the y-axis represents the feature of the data: the number of occurrences of the IP address. In order to display the result clearly, the data will be arranged on the x axis according to the order of their feature values.

- K-means clustering

The 'KMeans' package from 'sklearn.cluster' is used to apply k-means clustering. And the k value, which is the number of cluster, is determined by elbow plot.
As the Figure 3(a) and Figure 4(a) show, k=2 is the optimal value for destination IP, k=2 and k=3 are both acceptable for source IP. In this report, the displayed source IP clustering result adopts k =3. The clustering result of IP addresses is shown in Figure 3(b) and Figure 4(b).

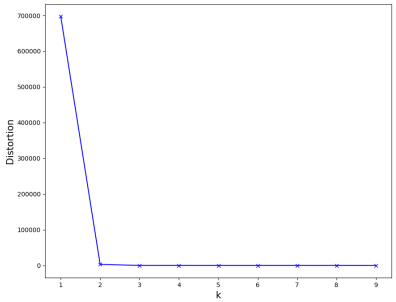


(a) The Elbow plot for the optimal k

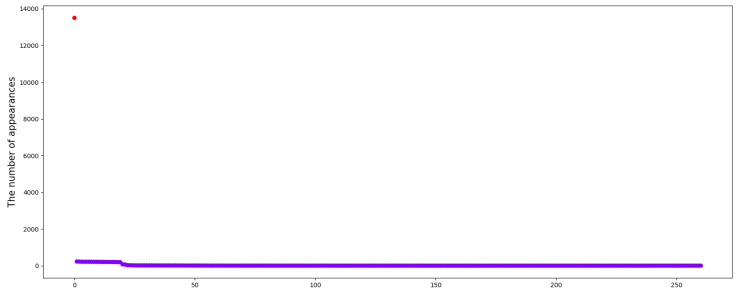


(b) K-means clustering result (k=3)

Figure 3: K-Means clustering for source IP addresses



(a) The Elbow plot for the optimal k

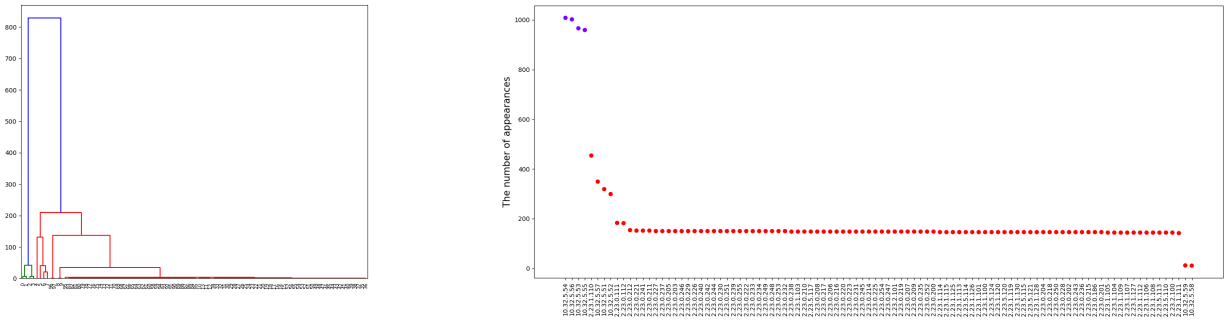


(b) K-means clustering result (k=2)

Figure 4: K-Means clustering for destination IP addresses

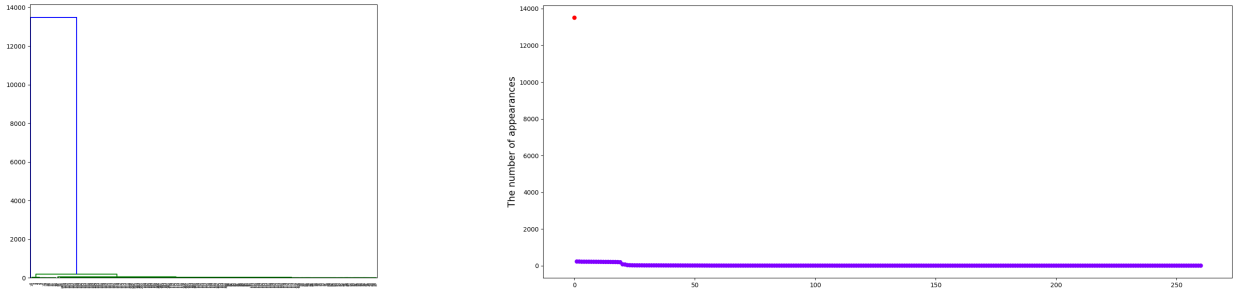
- Hierarchical Clustering

The 'scipy.cluster.hierarchy' pacakge is used to apply Hierarchical Clustering. The number of clusters can be determined by observing the hierarchical dendogram.
As the Figure 5(a) and Figure 6(a) show, for both source IP and destination IP, 2 is the best number of clusters. The clustering result of IP addresses is shown in Figure 5(b) and Figure 6(b).



(a) Hierarchical dendogram (b) Hierarchical clustering result

Figure 5: Hierarchical clustering for source IP addresses



(a) Hierarchical dendogram (b) Hierarchical clustering result

Figure 6: Hierarchical clustering for destination IP addresses

- Gaussian mixture models and EM algorithm

The 'GaussianMixture' pacakge from 'sklearn.mixture' is used to apply Gaussian mixture models (GMM) Clustering. I tried to use the BIC criterion to select the number of components in GMM by using calling the function 'GaussianMixture.bic', but the result of this method shows that the optimal number of clusters is more than 10, which is not ideal. Thus, in my program, the number of components in GMM is still 2 for both source IP and destination IP. Since the clustering result plot is similar to the previous one, it is not shown in the report.

Problem 4
Using 4 clusters for source IP addresses (split them at up to 20 records, 21-200,201-400, $>$ 400) and 4 clusters for destination IP addresses (split them at up to 40 records, 41-100, 101-400, $>$ 400), investigate the relation between source and destination. Can you determine conditional probabilities? Can you illustrate this graphically?

Solution: To find the relation between source and destination, we need to calculate the conditional probabilities. After calculating through the program, *P(Dest Cluster|Source Cluster)* is shown in Table 1, and *P(Source Cluster|Dest Cluster)* is shown in table 2.

Table 1: *P(Dest Cluster|Source Cluster)*

	d1	d2	d3	d4
s1	0.608	0.391	0	0
s2	0	0	0	1
s3	0.067	0.029	0.902	0
s4	0.172	0.022	0.701	0.103

Table 2: *P(Source Cluster|Dest Cluster)*

	s1	s2	s3	s4
d1	0.016	0	0.077	0.905
d2	0.065	0	0.211	0.722
d3	0	0	0.221	0.779
d4	0	0.966	0	0.033

We can visualize these data by stacked histograms as shown in Figure 8 and Figure 9:

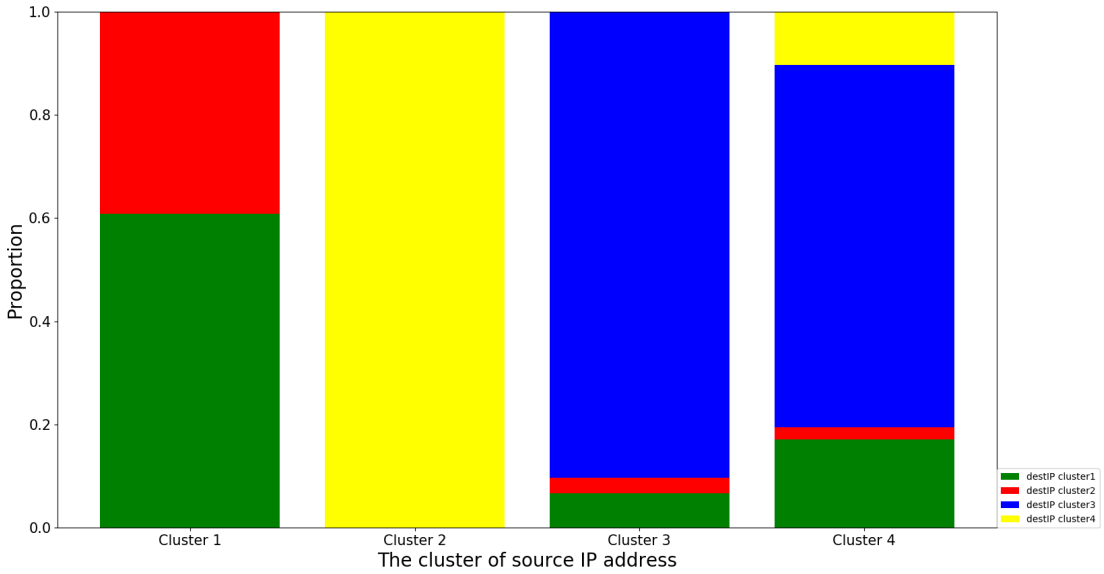


Figure 7: P(Destination Cluster|Source Cluster)

Figure 7 shows that source-cluster 2 always contact a destination in dest-cluster 4; All source IP in source-cluster 1 will go to the destination IP in dest-cluster 1 and 2; Most of the source IP in source-cluster 3 and 4 contact the destination in dest-cluster 3.

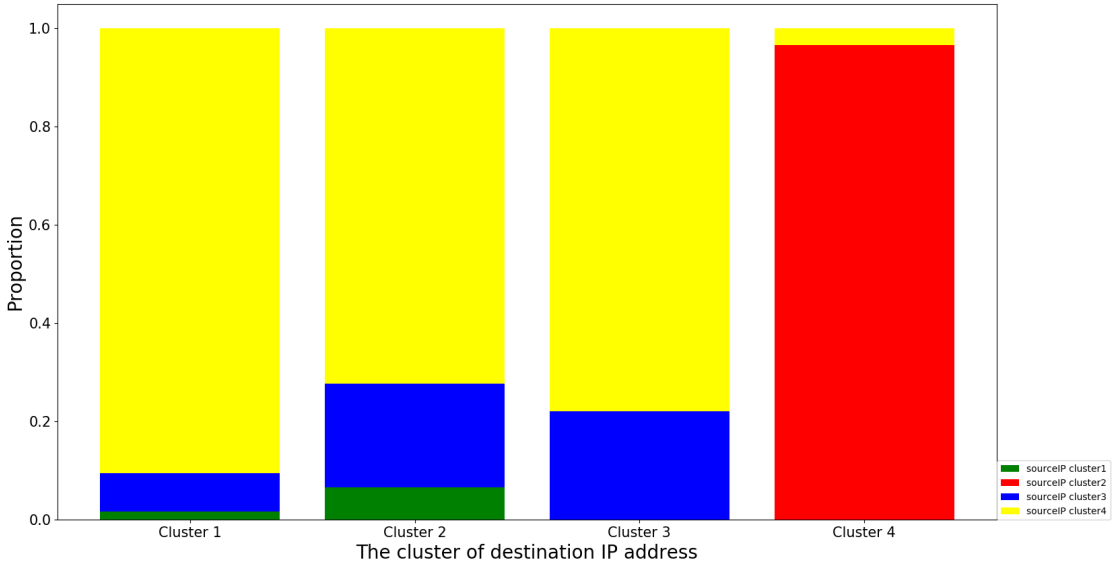


Figure 8: P(Source Cluster|Destination Cluster)

Figure 8 shows that most of the intrusions in dest-cluster 1,2 and 3 are from source-cluster 4; Almost all intrusions in dest-cluster 4 are from source-cluster 2.

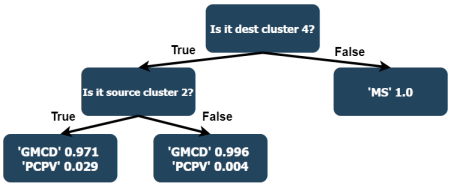
Problem 5
Write some of your own code to learn a decision tree using the 2 features above. Display your tree. In how many cases does your learnt decision tree give an unambiguous answer?

Solution: In this report, a binary tree using ID3 algorithm will be generated.
In order to make the decision tree have strong classification ability, the core question is: how to make the data be partitioned at each node in the best way? In another word, how to make the subsets data as pure as possible. In this report, the ID3 algorithm is adopted to build the decision tree, it uses information entropy to quantify the purity at each node. And the information gain calculated from the information entropy can quantify how much uncertainty can be reduced by a split method at a node. So my basic idea

of building a decision tree is: calculate the information gain of every possible partition condition at the nodes, and split the dataset at each node by the partition condition that can get the highest information gain until the node cannot be partitioned.
The outline of my decision tree program is: There is a function named "build_tree", the input is the training dataset. It will call the "determine_node" function to determine the best split condition and gain of the current node. If the gain equal to zero, the leaf will be added. Otherwise the program will according to the determined split condition split the dataset into two subsets: one subset meets the node's condition, the other one does not. The subsets then become the input of two child nodes, and the child nodes will call the "build_tree" function again so that the program can recursively build up the tree from the root node to leaves.
After training the decision tree, we can use the tree to classify data. However, not every data can be classified unambiguously, so on each leaf, the program will show all possible classification results and the probability of each result.
The learnt tree is displayed in Figure 9. Figure 9(a) is generated by program and Figure 9(b) is made by other software for clearer display. This decision tree has two nodes and three leaves. The tree is tested by testing set, the average accuracy is around 0.98.

```
dest_cluster == dest_cluster4?
--> True:
  source_cluster == source_cluster2?
  --> True:
    Predict {' Generic Protocol Command Decode': 0.971, ' Potential Corporate Privacy Violation': 0.029}
  --> False:
    Predict {' Generic Protocol Command Decode': 0.996, ' Potential Corporate Privacy Violation': 0.004}
--> False:
  Predict {' Misc activity': 1.0}
0.9837221920781335
```

(a)



(b)

Figure 9: Decision Tree of Question 5

The leaves on the decision tree represent the possible classification result. For example, the classification result of the data:(source-cluster 2, dest-cluster 4) can be " 0.971 probability to be *Generic Protocol Command* and 0.029 probability to be *Potential Corporate Privacy Violation*". Thus, to find the data that the decision tree can make unambiguous classification, we can compare the max probability value with the threshold value that constitute "quite certain". Because in this question all of the max probability is larger than 0.97, so we can accept '=1.0' as unambiguous answer. The unambiguous set when classification confidence = 1.0 is shown in the Figure 10, the number of cases is 8 the classification of all of them is 'Misc activity'.

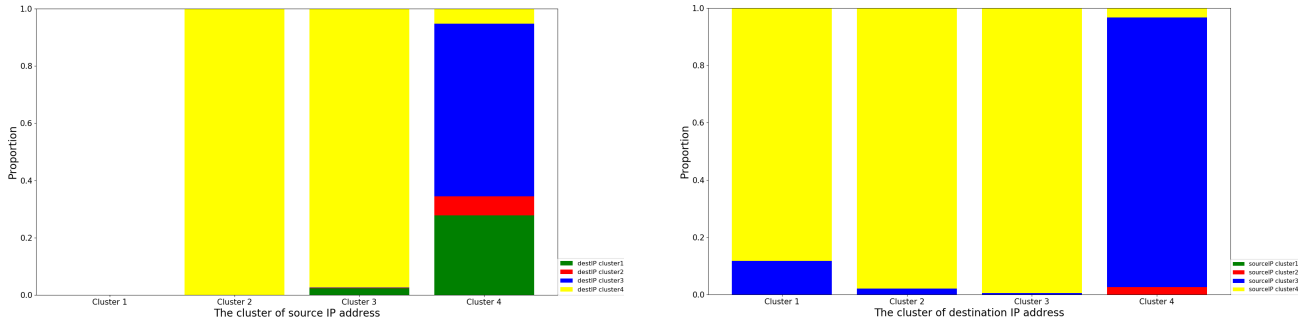
```
['source_cluster4', 'dest_cluster1', ' Misc activity']
['source_cluster1', 'dest_cluster2', ' Misc activity']
['source_cluster4', 'dest_cluster3', ' Misc activity']
['source_cluster3', 'dest_cluster3', ' Misc activity']
['source_cluster3', 'dest_cluster1', ' Misc activity']
['source_cluster4', 'dest_cluster2', ' Misc activity']
['source_cluster3', 'dest_cluster2', ' Misc activity']
['source_cluster1', 'dest_cluster1', ' Misc activity']
```

Figure 10: The unambiguous set when classification confidence = 1.0

Problem 6
Examine the dataset coursework2.csv. Using the same clusters of IP addresses, are the patterns observed in Q4 still valid? How about the decision tree in Q5?

Solution: Process and clustering the data from "coursework2.csv" with the same standards as Q4. Figure

11 shows the conditional probabilities. Compared with Q4, the probability distribution of the two is very different.
From Figure 11(a), we can find that there is no data belonging to source cluster 1 in the dataset; And most of the source IP in source-cluster 3 contact the destination in dest-cluster 4 in "coursework2.csv". But still same as Q4, source-cluster 2 always contact a destination in dest-cluster 4.
From Figure 11(b), we can find that it is similar to Q4, most of the intrusions in dest-cluster 1,2 and 3 are from source-cluster 4. However, unlike Q4, in dest cluster 4, most of the invasions is from source cluster 3.



(a) P(Destination Cluster|Source Cluster)

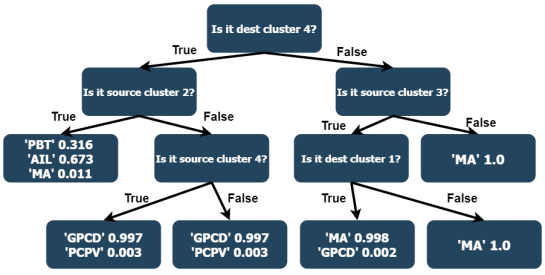
(b) P(Source Cluster|Destination Cluster)

Figure 11: Conditional Probability

Compared with "coursework1.csv", the classification in "coursework2.csv" is more abundant. The latter has two new classifications that the former does not:*"Potentially Bad Traffic"* and *"Attempted Information Leak"*. Thus, the decision tree in Q6 has more nodes and leaves than Q5. Decision tree trained by "coursework2.csv" is shown in Figure 12, it has 5 nodes and 6 leaves. It is worth noting that the two leaves under the node *"Is it source cluster 4"* actually have slight differences.

```
dest_cluster == dest_cluster4?
--> True:
  source_cluster == source_cluster2?
  --> True:
    Predict {' Potentially Bad Traffic': 0.316, ' Attempted Information Leak': 0.673, ' Misc activity': 0.011}
  --> False:
    source_cluster == source_cluster4?
    --> True:
      Predict {' Generic Protocol Command Decode': 0.997, ' Potential Corporate Privacy Violation': 0.003}
    --> False:
      Predict {' Generic Protocol Command Decode': 0.997, ' Potential Corporate Privacy Violation': 0.003}
--> False:
  source_cluster == source_cluster3?
  --> True:
    dest_cluster == dest_cluster1?
    --> True:
      Predict {' Misc activity': 0.998, ' Generic Protocol Command Decode': 0.002}
    --> False:
      Predict {' Misc activity': 1.0}
  --> False:
    Predict {' Misc activity': 1.0}
```

(a)



(b)

Figure 12: Decision Tree of Question 5

Same as Q5we set the classification confidence equal to 1.0. In this situation, there are only 5 cases as the Fig13 shown, the classification of all of them is also "Misc Activity". The comparison between Q5 and Q6 with different classification confidence equal to 0.95 shown in Appendix.

```
['source_cluster4', 'dest_cluster1', ' Misc activity']
['source_cluster4', 'dest_cluster3', ' Misc activity']
['source_cluster4', 'dest_cluster2', ' Misc activity']
['source_cluster3', 'dest_cluster3', ' Misc activity']
['source_cluster3', 'dest_cluster2', ' Misc activity']
```

Figure 13: The unambiguous set when classification confidence = 1.0

Appendix

```
['source_cluster2', 'dest_cluster4', ' Generic Protocol Command Decode ...
['source_cluster4', 'dest_cluster3', ' Misc activity']
['source_cluster3', 'dest_cluster1', ' Misc activity']
['source_cluster4', 'dest_cluster4', ' Generic Protocol Command Decode ...
['source_cluster2', 'dest_cluster4', ' Potential Corporate Privacy Vio ...
['source_cluster4', 'dest_cluster1', ' Misc activity']
['source_cluster3', 'dest_cluster3', ' Misc activity']
['source_cluster4', 'dest_cluster2', ' Misc activity']
['source_cluster3', 'dest_cluster2', ' Misc activity']
['source_cluster1', 'dest_cluster2', ' Misc activity']
['source_cluster1', 'dest_cluster1', ' Misc activity']
['source_cluster4', 'dest_cluster4', ' Potential Corporate Privacy Vio ...
```

(a) Q5:coursework1.csv

```
['source_cluster3', 'dest_cluster4', ' Generic Protocol Command Decode ...
['source_cluster4', 'dest_cluster2', ' Misc activity']
['source_cluster4', 'dest_cluster3', ' Misc activity']
['source_cluster4', 'dest_cluster1', ' Misc activity']
['source_cluster4', 'dest_cluster4', ' Generic Protocol Command Decode ...
['source_cluster3', 'dest_cluster1', ' Misc activity']
['source_cluster3', 'dest_cluster4', ' Potential Corporate Privacy Vio ...
['source_cluster3', 'dest_cluster2', ' Misc activity']
['source_cluster3', 'dest_cluster3', ' Misc activity']
['source_cluster3', 'dest_cluster1', ' Generic Protocol Command Decode ...
['source_cluster4', 'dest_cluster4', ' Potential Corporate Privacy Vio ...
```

(b) Q6:coursework2.csv

Figure 14: The unambiguous set when classification confidence = 0.95