Data analytics for Cyber security

-Anomaly Detection-

Vandana P. Janeja

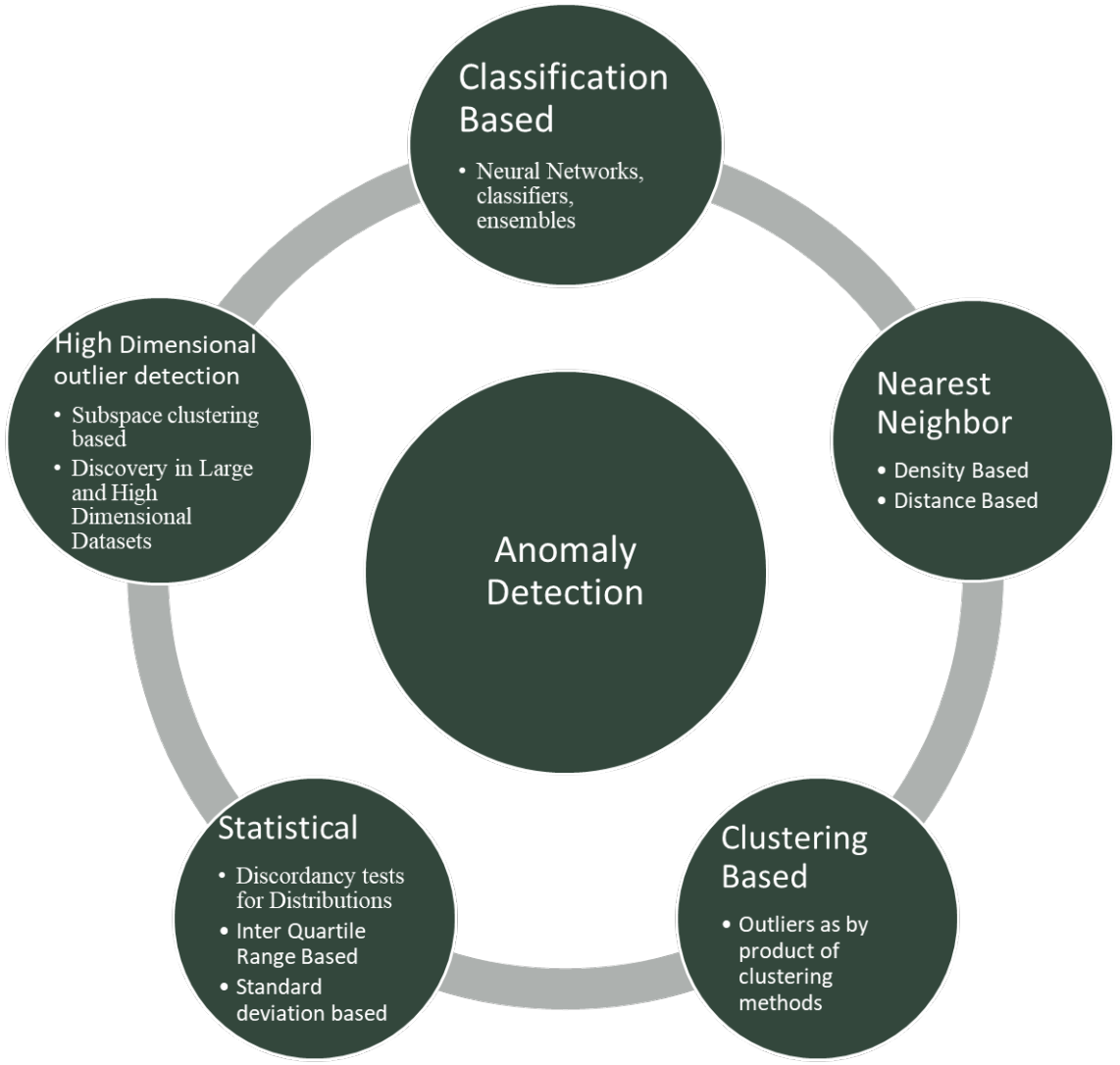
Outline

Anomaly Detection Methods

- Statistical Outlier Detection Tests
- IQR-Based Outlier Detection
- Density-Based Outlier Detection
  OPTICS-OF Identifying Local Outliers
- Distance-Based Outlier Detection
- Outlier Detection through Clustering

Anomaly Detection

- Anomalies or outliers are objects or groups of objects, which are deviant with respect to other normal objects
- Anomalous behavior can be caused by extreme values in some dimensions, which could be an inherent rare property of the anomalous objects
- In high dimensional data this rarity increases with the increase in the number of dimensions
- Data itself becomes sparse in higher dimensions; this will further aggravate the discovery of outliers
- Other anomalous situations may be caused due to measurement errors or phenomena which are causing the deviations



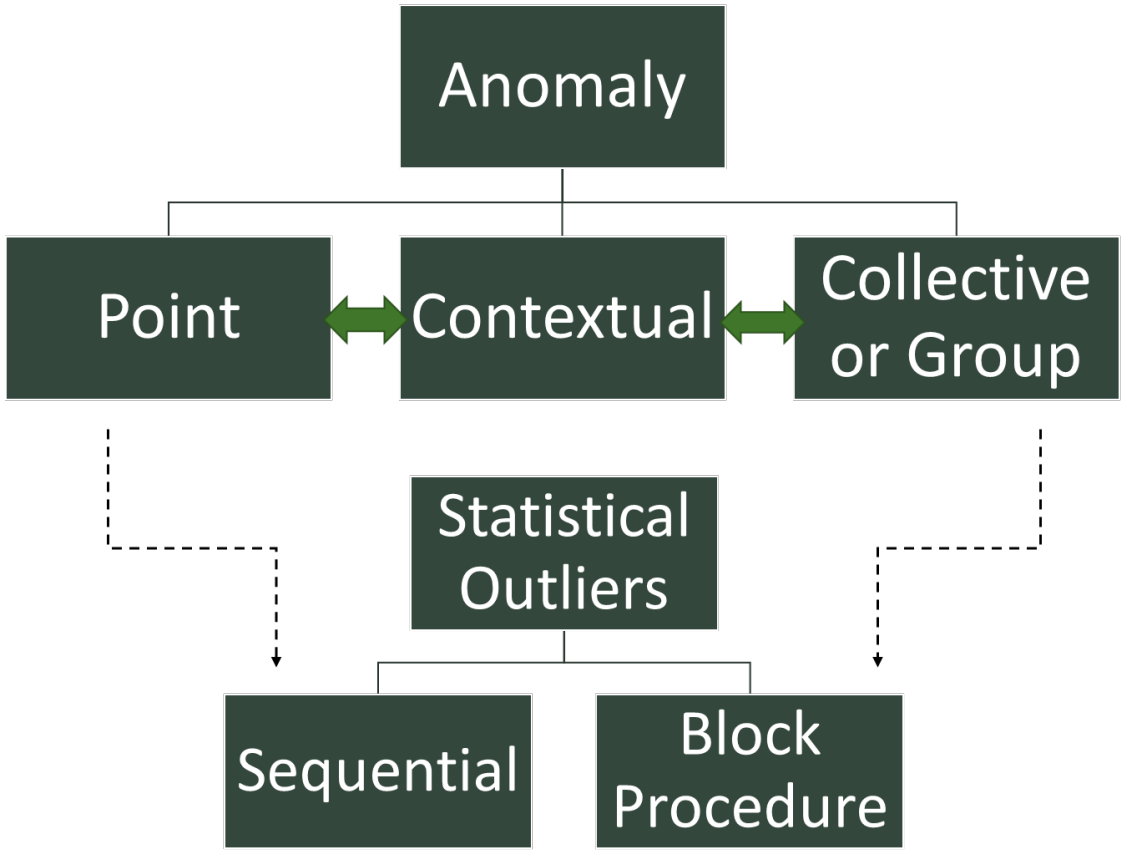
Anomaly Detection Methods

- Outlier detection techniques have been adopted from the statistical literature and treat the outliers as parametric outlier detection, in which the data is assumed to fit a certain distribution
- The tail region of the distribution, such as in a normal distribution, would constitute the outliers; However, in real world datasets, the data does not fit into a standard statistical distribution
- Other techniques detect outliers using metrics such as distance, distance from nearest neighbors, density, deviation and other such measures
- By-product of clustering methods: The data is first clustered using some clustering technique and the data points, which lie outside of a cluster, are considered to be outliers.
- In some cases part or whole of a cluster of data points could be an outlier based on the problem being studied
- Classification based techniques identify and predict anomalies based on historic anomaly signatures

Statistical Outlier Detection Tests

- The distribution based parametric methods are discussed in the statistical outlier detection literature which assume that data fits into a pre specified distribution, such as normal distribution and then identifies the outliers with respect to the model using Discordancy Tests
- The discordancy tests are developed based on the properties of the distribution
- Discordancy tests assumes a standard distribution of the data set and if the discordancy test finds the outliers then it tries to fit the data to the alternative distribution, however the underlying assumption is that the data fits a distribution, which might not be the case always, especially for multidimensional real world data
- If there is a normal distribution where x_1 and x_n are the extremes, a discordancy test will check if x_n apart from being an extreme is also statistically unreasonable, if it is then it is called a discordant upper outlier
- The followup step would be to show that the outlier comes from the alternative distribution and not the working distribution, which is a difficult task in itself
- A key metric is to see the fit of the data point using measures such as deviation/spread statistics such that: $d = \frac{\mu - x1}{s}$ where μ is a measure of central tendency in this case the mean and s can be a known measure of spread such as range
- For example, if d <= 1 then x_1 is not an outlier, on the other hand if d >1 it is an outlier

Statistical Outlier Detection Tests



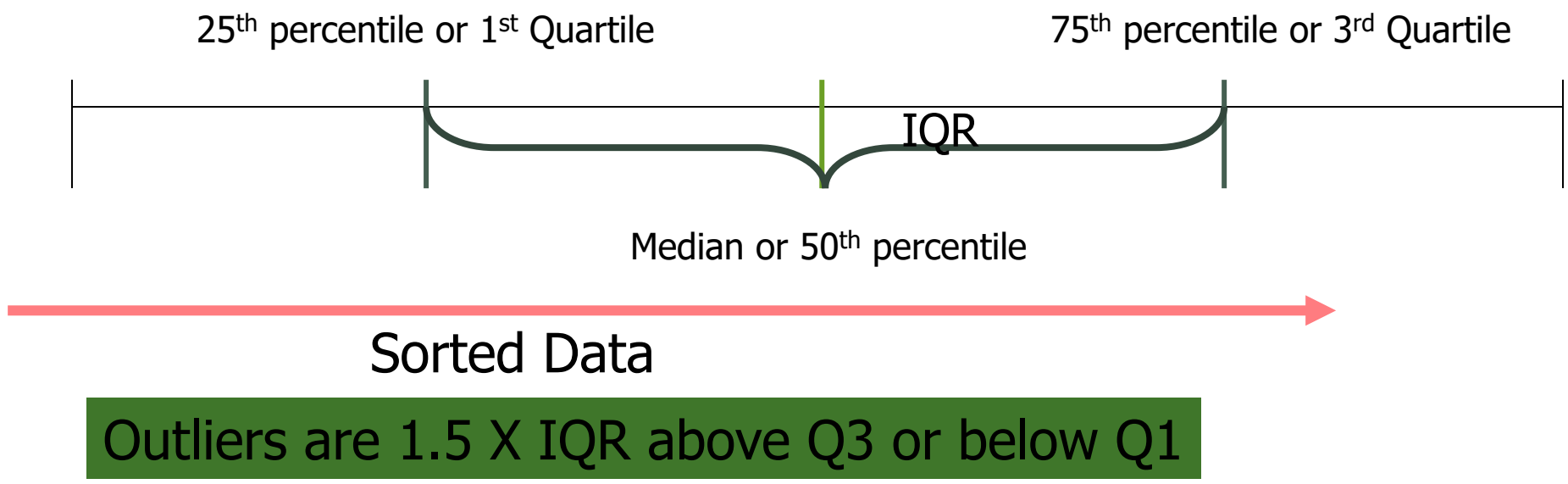
- There are two major procedures for outlier detection in statistical studies Block Procedures and Consecutive (Or Sequential) Procedures
- *Consecutive (Or Sequential) Procedures*
  - sample size is not fixed, but determined in each cycle in relation to the values of earlier observations
  - Each observation is screened for discordancy in relation to the already screened set of observations
  - This procedure can be done in two ways: outside inwards which means that the extreme outlier is tested first then second most extreme and so on, but more widely preferred way of doing it is the other way round (starting with the least likely element to be an outlier)
- *Block Procedures*: two or more observations are tested together for outlierness
- Consecutive procedures can suffer from challenges such as Masking and Swamping (Barnett and Lewis 1994).
- *Masking* is an erroneous judgment arising out of a single outlier test. Suppose if through a discordancy test for x_n it is declared and outlier with respect to all the n-1 values then x_{n-1} is further tested, but if x_n is not declared an outlier with respect to all n-1 values then x_{n-1} is not tested
- This could be a problem in a scenario where x_n is not an outlier with respect to the adjacent value of x_{n-1}, but both x_n and x_{n-1} are outliers when tested together with respect to all n-2 values.
- Let us consider a small sample values: 8, 10, 11, 13, 1000, and 1005
- If tested individually 1009 will not be declared as an outlier since it is close to 1000 but if 1000 and 1005 are tested together against other values then they are outliers with respect to all other n-2 values
- Swamping on the other hand results from using a block or consecutive procedure to test for discordancy i.e. testing two values together
- For example in a sample: 5,8,10,12,1000
- If two values are tested together then they are considered outliers but value 12 is not an outlier, so in essence 12 is swamped by 1000
- These procedures have similarity to the point anomaly detection and collective anomaly detection discussed in the data mining literature
- This behavior is also seen in majority of outlier detection techniques when the incorrect context for comparison can lead to missing outlier values

IQR-Based Outlier Detection

- Inter quartile range based outlier detection aims to identify the range of the data and find extreme values which are far away from the range
  - First, data is sorted in increasing order
  - The median of the data is identified which becomes the second quartile
  - All the numbers to the left are smaller than the median and all the numbers to the right are greater
  - The median of these left and right hand sides are identified as first quartile and third quartile respectively
  - The data is now divided into four chunks of data ordered by the quartiles
  - The difference between the third quartile and the first quartile is the Inter Quartile Range (IQR)
  - IQR based outlier detection assumes that the IQR should have the majority of the data distribution measuring the spread of the data
  - Now if this IQR range is further stretched beyond the first and third quartile such that the lower threshold is first quartile minus IQR times 1.5 and the upper threshold is third quartile plus the IQR times 1.5
  - Points which fall beyond these thresholds at the lower extreme or the upper extreme are the outliers
- IQR is primarily a univariate outlier detection method

Interquartile based outliers- boxplot

K percentile, k% of the data entries lie at or below kth value



Outliers are 1.5 X IQR above Q3 or below Q1

Example: Interquartile based outliers

Observations	packet size		
1	7.012472	Outlier	
2	8.771873		
3	9.360618		
4	9.428184		
5	9.456278	9.476627769	**Q1**
6	9.496977		
7	9.608662		
8	9.881874		
9	9.922703		
10	10.11899		
11	10.47658	median	**Q2**
12	10.51159		
13	10.63797		
14	10.91105		
15	10.98602		
16	11.06	11.08674809	**Q3**
17	11.1135		
18	11.48358		
19	11.69379		
20	11.95874		
21	25	Outlier	
IQR	1.61012		
IQR*1.5	2.41518		
Thresholds	7.061447	**lower threshold**	
	13.50193	**upper threshold**	

Density-Based Outlier Detection: OPTICS-OF Identifying Local Outliers

- OPTICS-OF is an extension of the OPTICS density based clustering algorithm
- It discovers outliers relative to their neighborhood
- Distance based outliers are good in a global sense, but when different densities of objects exists within the data, distance based discovery is not adequate
- Each outlier is assigned with a degree by which the object is an outlier
- Outliers, relative to their local surrounding space, motivates a formal definition of local outliers
- An object is mostly defined as an outlier or not an outlier; however there could be a degree of outlierness attached to each object based on the local density where position of each object affects the other
- The basic concepts of ε-neighborhood, K-distance are adapted from the algorithm DBSCAN and the concepts of core-distance and reachability distance from the algorithm OPTICS
- Optics – OF defines a concept called, Local Reachability which is measured in terms of Local reachability density of point p which is given as: Lrd(p)=1/average reachability distance of minpts nearest neighbors of P

OPTICS-OF Example

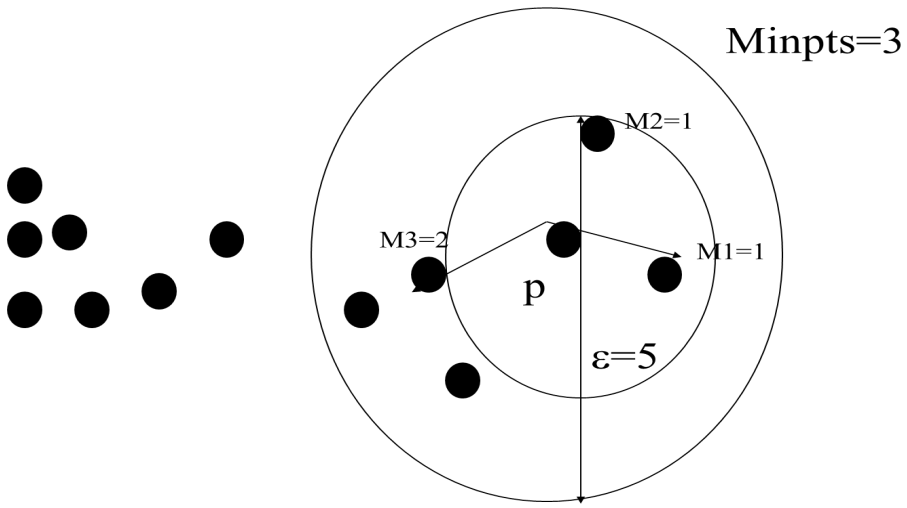
- Min pts are 3 although there are more than 3 pts in the neighborhood
- Only reachability distance to the 3 points needs to be measured
- Thus, Lrd(p)=1/((1+1+2)/3) which is computed as =>1/(4/3)=3/4= 0.75
- Outlier Factor of p is measured in terms of the uniformity in density of p's neighborhood, thus, it takes into consideration the local reachability density of all the min points' nearest neighbors of p, in the above example these are M1, M2, M3
- The outlier Factor of p is: $OF(p) = \frac{\frac{\sum lrd\,(o)}{lrd(p)}}{N}$
- So the Outlier factor is: (0.75+1+0.5/.75)/3=(2.25/.75)/3=1
-
- The OF =1 represents a cluster of Uniform density
- The lower is p's Lrd and the higher is the Lrd of p's MinPts nearest neighbors, the higher will be p's Outlier Factor

OPTICS-OF Example

$$lrd(p) = \frac{1}{\text{average reachability distance of minpts Nearest neighbors of P}}$$

$$lrd(p) = \frac{1}{((1+1+2)/3)}$$

=>1/(4/3)=3/4= 0.75



$$OF(p) = \frac{\frac{\sum lrd\ (o)}{lrd(p)}}{N}$$

$$lrd(M1) = \frac{1}{((1+1+2)/3)} = 0.75$$

$$lrd(M2) = \frac{1}{((1+1+1)/3)} = 1$$

$$lrd(M1) = \frac{1}{((2+2+2)/3)} = 0.5$$

So the Outlier factor is:
(0.75+1+0.5/.75)/3=1

Distance-Based Outlier Detection

- Distance based outlier detection algorithms continues the idea of a unified notion of mining outliers, which extends on the statistical based outlier detection
- It generalizes the notion of outliers provided by many of the discordancy tests for known distributions, like Normal, Poisson, Binomial etc.
- Although it proposes a partitioning based technique for distance based outlier detection, it also elaborates on how this approach can unify the various discordancy tests available in the statistical/distribution based outlier detection techniques
- The time complexity is linear with respect to number of objects and the algorithms are efficient for low dimensional datasets
- The key idea is that outliers are far away from their neighbors in terms of a distance threshold

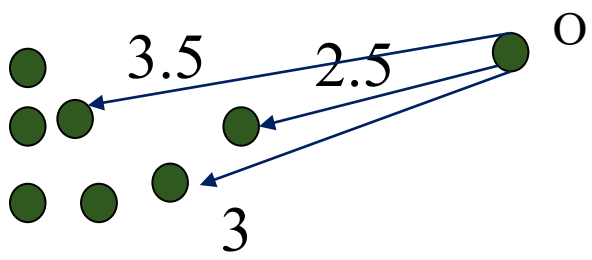
Distance-Based Outlier Detection

- The distance based outliers are identified with two parameters p, D namely DB (p, D)
- An object O in a dataset T is a DB (p, D) outlier if at least a fraction p of the objects in T are at a greater distance D from O
- Outliers are objects, which do not have enough neighbors within a certain distance
- This eliminates the computation for fitting the observed distribution into a standard distribution and in selecting discordancy tests
- For many discordancy tests for standard distributions if an object is shown to be an outlier then it is also a DB (p, d) outlier
- Parameters p, D are set by a user in case of non-standard distributions, this is a trial and error approach, where the user interactively changes the p, D variables to evaluate the outcome.
- This approach identifies outliers based on the distance of a candidate outlier from a set of neighboring points
- it is not dependent on clustering structure of the dataset
- The outliers detected however are dependent on user input
- Most of these algorithms are sensitive to the user input like the distance threshold, or the value of p to identify the number of points for comparison to the candidate point
- This approach is not suitable for high dimensional data as the Euclidean distance metric being used does not produce accurate results over high dimensions
- Distance metric is a point metric, therefore it detects outliers which are at a distance greater than a parameter, and it does not take into consideration outliers which could be as a result of a very small distance measure
- This technique is mainly geared towards detecting global outliers.

Distance-Based Outlier Detection - Example

The distance based outliers are identified with two parameters p, D namely DB (p, D). An object O in a dataset T is a DB (p, D) outlier if at least a fraction p of the objects in T are at a greater distance D from O.

p=3
D=2



The fraction of points p is 3 and distance threshold is 2
The point O is at least greater than distance 2 from at least 3 points and can be labelled an outlier

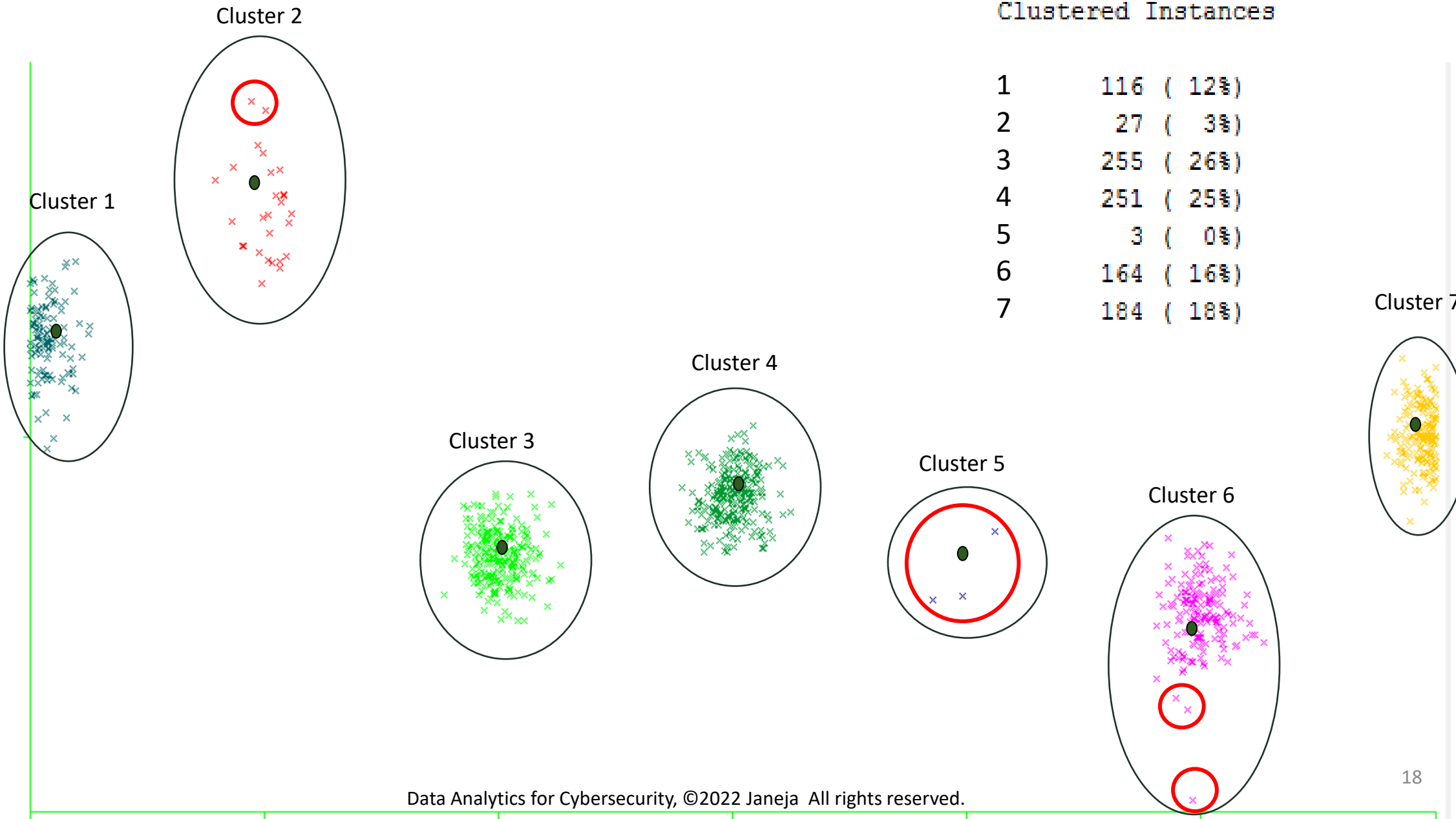
Distance-Based Outlier Detection : Challenges

- Distance based outlier detection methods suffer from the curse of dimensionality
- In higher dimensions data becomes sparse, points tend to become equidistant
- This adversely affects the methods based on spatial density, unless data follows certain simple distributions
- For outlier detection a subset of the entire attribute set should be enough to detect or indicate the outliers
- If we consider cross sections of the data in various dimensions we can find outliers

Outlier Detection through Clustering

- Some clustering algorithms such as DBSCAN identify outliers as a byproduct of clustering
- Majority of clustering algorithms have to be post processed to discover outliers
- Outliers can be seen as a byproduct of the clustering process
- In traditional clustering such as K-Means, by the end of the cluster analysis, we assign each data point to one of the clusters, so basically there are no unassigned data points
- An example of clustering the SANS data for the last 1000 days records, targets and sources
  - This clustering was generated using Weka, Simple Kmeans
  - Since all the data is assigned to clusters, does that mean that there are no outliers in the data?
  - If there are still outliers, how do we filter out the outliers?
  - There are a few possibilities in a clustering output: (a) a point may be placed in a cluster by itself. (b) A data point may be part of a cluster but far from the centroid. (c) A small cluster with a lot fewer points than other clusters can be far from the other clusters. (d) In DBScan type clustering points will not be connected or in a cluster but a distinct outlier.
- If a point is placed in its own cluster or is identified by the algorithm as an outlier such as in the case (a) and (d) then clearly a point anomaly has been detected. In the other two cases special analysis will need to be performed to identify anomalies

Outlier Detection through Clustering: Example



Outlier Detection through Clustering: Example

- *Small clusters:*
  - If the data is clustered into groups of different density and if there is a cluster with fewer points than the other clusters this cluster can be a candidate collective anomaly
  - In this case the density of the candidate cluster is highly deviant than the other clusters
  - There can be other measures such as the distance between candidate cluster and non-candidate cluster and if candidate cluster is far from all other non-candidate points, they are potential outliers
- *Larger clusters with scattered points:*
  - There may be bigger clusters where the points in the cluster are not tightly bound around the centroid
  - In this case the distance of every point can be measured to the centroid and the points farthest from the centroid can be identified
  - This is similar to a distance based or a k-Nearest Neighbors based outlier detection inside such clusters
  - These clusters can be identified by looking at the Sum of Squared distance of the clusters and the ones with a high SSE can be candidates for post process through distance based outliers

References

- V. Barnett and R. Lewis. Outliers in statistical data. John Wiley and Sons, 1994
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.
- Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12), 3448-3470.
- Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. computers & security, 28(1), 18-28.
- Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2014). Network anomaly detection: methods, systems and tools. Ieee communications surveys & tutorials, 16(1), 303-336.
- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (1999, September). Optics-of: Identifying local outliers. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 262-270). Springer Berlin Heidelberg.
- Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. The VLDB Journal—The International Journal on Very Large Data Bases, 8(3-4), 237-253.
- Sugiyama, M., & Borgwardt, K. (2013). Rapid distance-based outlier detection via sampling. In Advances in Neural Information Processing Systems (pp. 467-475).
- Cao, L., Yang, D., Wang, Q., Yu, Y., Wang, J., & Rundensteiner, E. A. (2014, March). Scalable distance-based outlier detection over high-volume data streams. In Data Engineering (ICDE), 2014 IEEE 30th International Conference on (pp. 76-87). IEEE.
- Zhang, K., Hutter, M., & Jin, H. (2009). A new local distance-based outlier detection approach for scattered real-world data. Advances in knowledge discovery and data mining, 813-822.
- Ramaswamy, S., Rastogi, R., & Shim, K. (2000, May). Efficient algorithms for mining outliers from large data sets. In ACM Sigmod Record (Vol. 29, No. 2, pp. 427-438). ACM.
- Aggarwal, C. C., & Yu, P. S. (2001, May). Outlier detection for high dimensional data. In ACM Sigmod Record (Vol. 30, No. 2, pp. 37-46). ACM.