Relative Outlier Cluster Factor(ROCF)

Relative Outlier Cluster Factor(ROCF)

ROCF Process

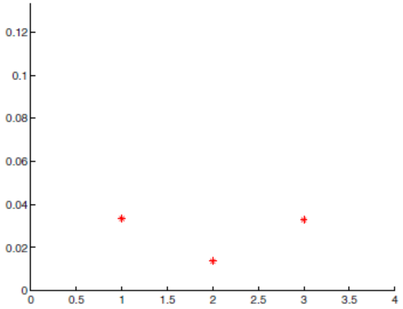
- 먼저 MkNN을 수행해 clustering과 scatter outlier detection을 하고, transition level(TL)와 $ROCF$를 정의하고 각 cluste에 대해 $ROCF$를 계산한다.
- Cluster들을 크기 기준 오름차순으로 정렬한 후, b번째 cluster의 ROCF 값이 0.1 보다 크면, 1부터 b번째까지의 cluster는 outlier cluster로 판단한다.
- 0.1보다 큰 ROCF를 가지는 cluster가 없으면, MkNN으로 감지한 scatter outlier만을 outlier로 판단한다.

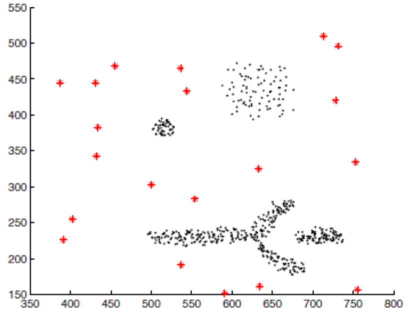
$$TL(C_i) = \frac{|C_{i+1}|}{|C_i|}, i = 1, 2, \ldots, n-1$$

$$ROCF(C_i) = 1 - e^{-\frac{TL(C_i)}{|C_i|}} = 1 - e^{-\frac{|C_{i+1}|}{|C_i|^2}}, \quad i = 1, 2, \ldots, n-1$$

$$|C_1| \le |C_2| \le \ldots \le |C_n|$$

$$max\{ROCF(C_i)\} \text{ and } ROCF(C_b) > 0.1, \text{ then } C_1, C_2, \ldots, C_b$$



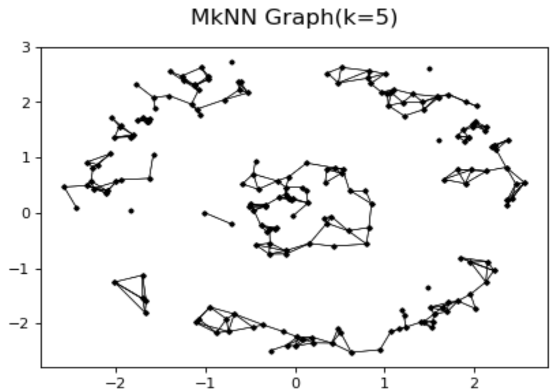


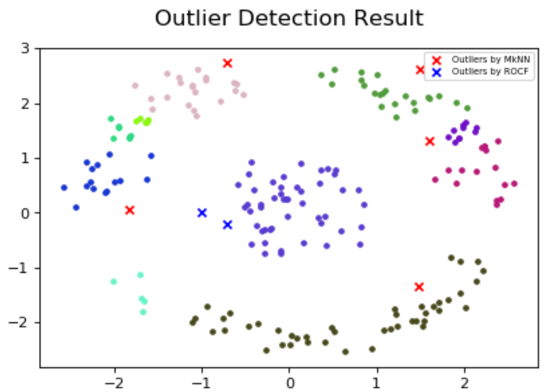
Relative Outlier Cluster Factor(ROCF)

MkNN(Mutual k-Nearest Neighbors)

- 기존에 구현을 계획한 Hu, Zhen. (2012)의 MkNN 방법론은 구현의 어려움과 함께 일반적인 outlier detection에 적용되기 어렵다는 한계가 있다.
- 보다 널리 이용되며 일반화하기 용이한 Maier et al. (2009)의 MkNN 방법론에 기초해, ROCF를 적용하기 전 rough clustering 및 scatter outlier detection 을 진행했다.
- 두 data point에 있어, $data\ j$가 $data\ i$의 k-distance 범위에 있고 $data\ i$ 또한 $data\ j$의 k-distance 범위에 있으면, 두 data point가 연결된 것으로 정의한다. 각 data point의 neighbors는 k 개로 제한된다.
- 각 data point는 mutual한 관계를 가지고 연결되므로, clustering 결과는 그래프로도 표현될 수 있다.

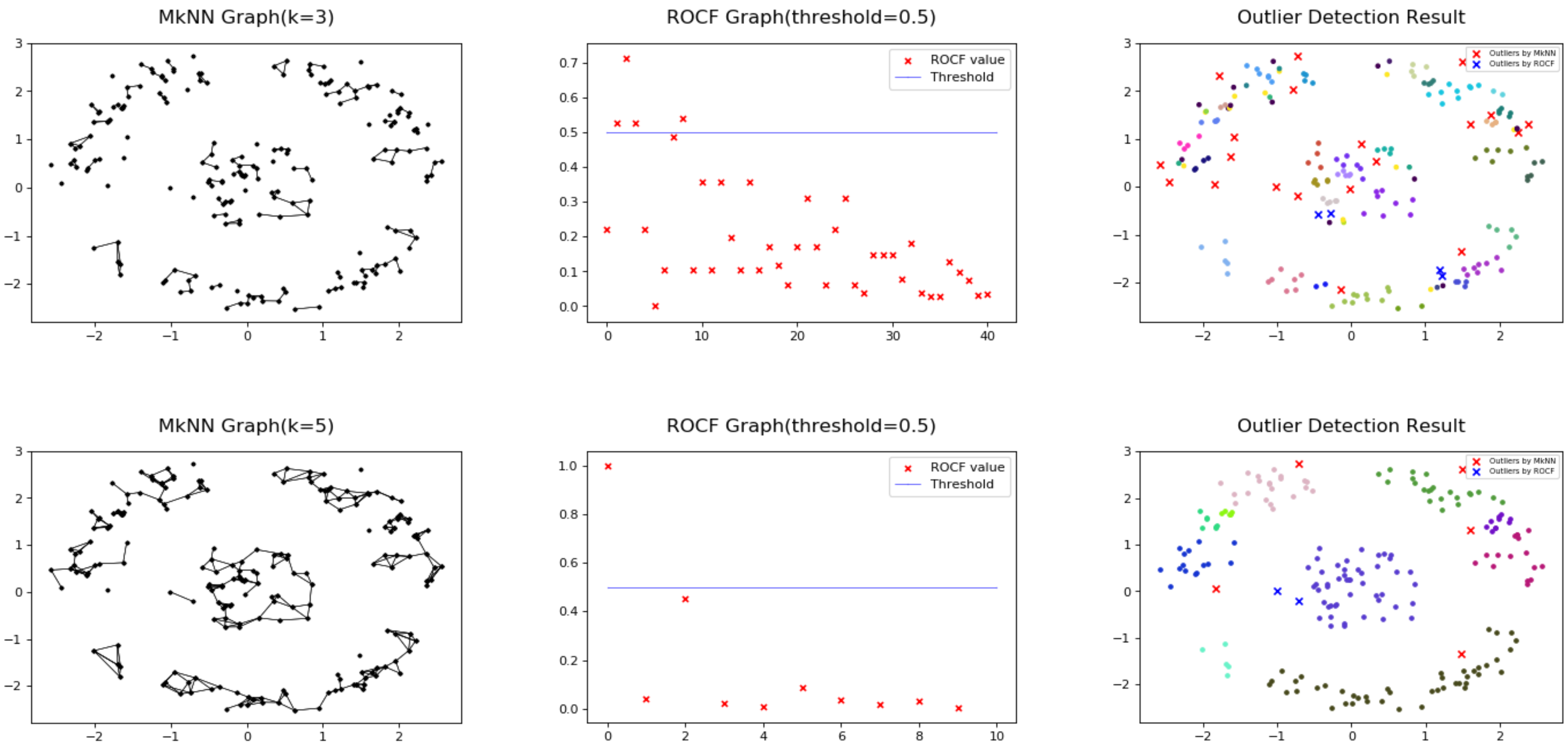
- *mutual k-nearest-neighbor graph* $G_{\mathrm{mut}}(n, k)$*:*
  X_i and X_j are connected if $X_i \in \mathrm{kNN}(X_j)$ and $X_j \in \mathrm{kNN}(X_i)$.





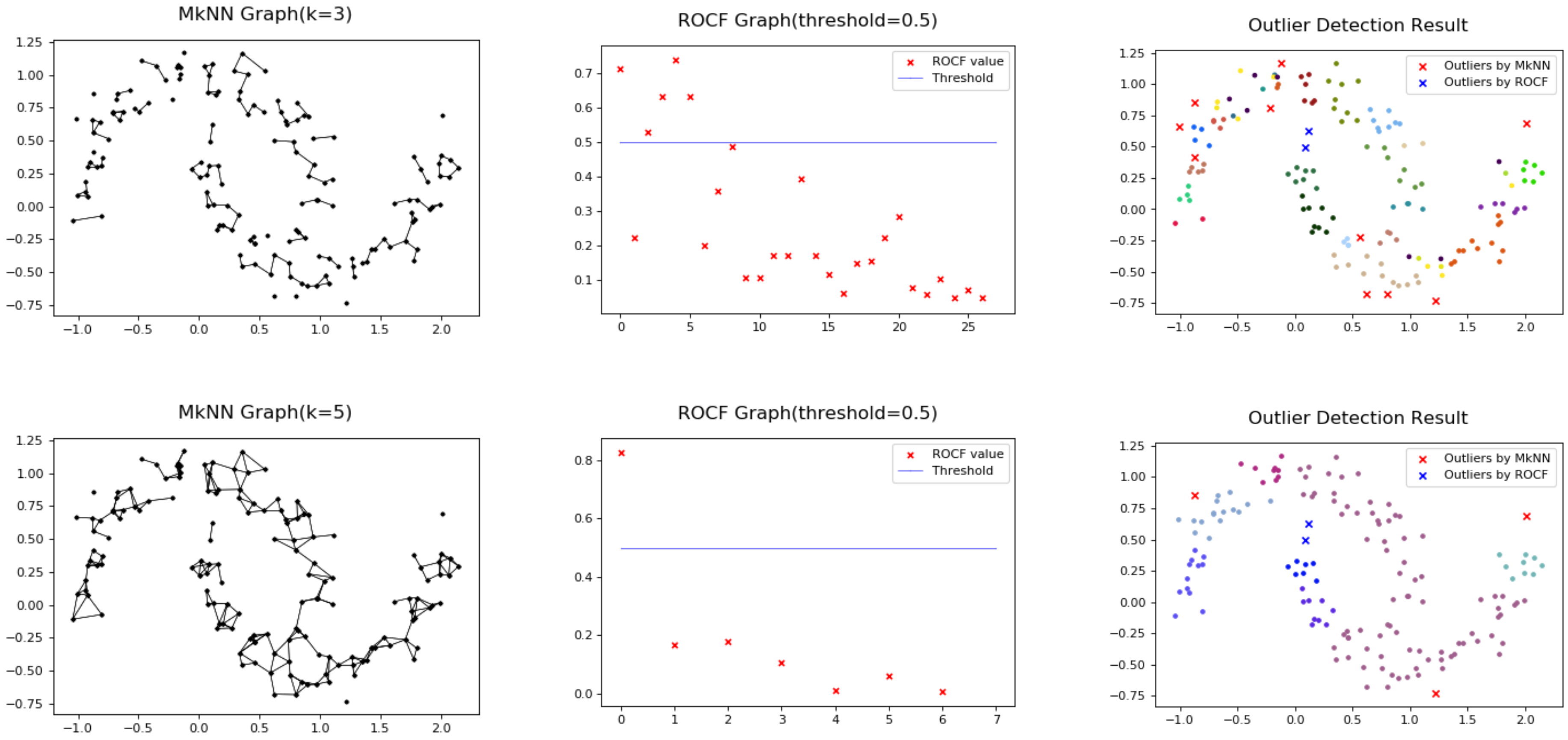
Relative Outlier Cluster Factor(ROCF)

- Ring dataset



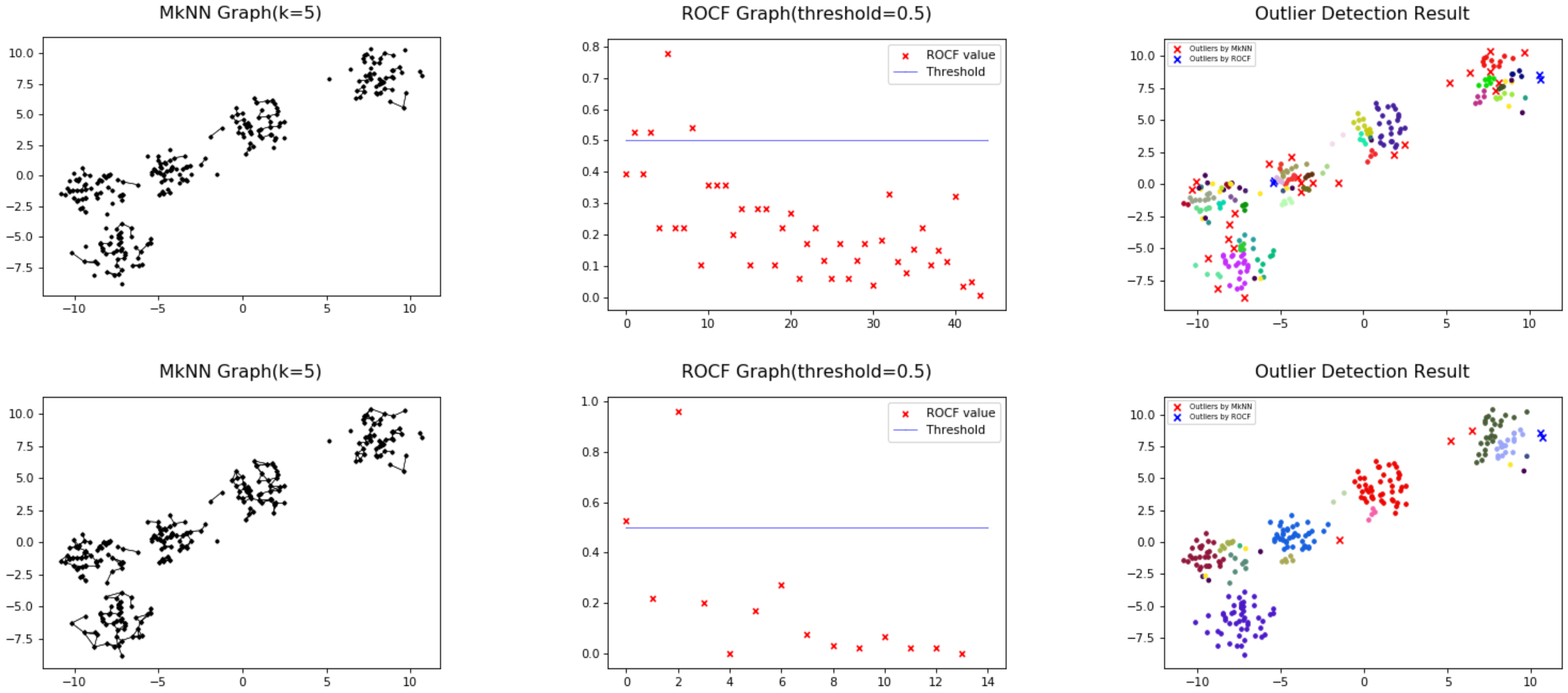
Relative Outlier Cluster Factor(ROCF)

- Moon dataset(n_samples= 150, noise= 0.1)



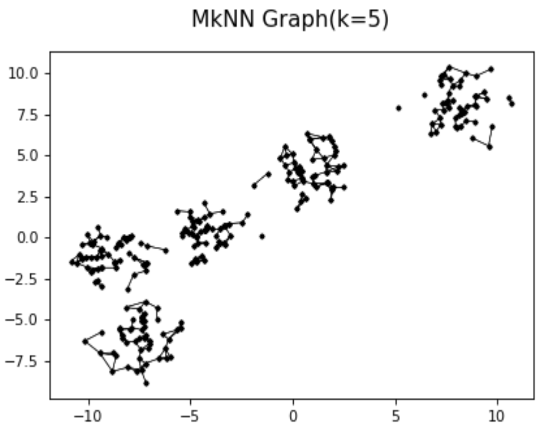
Relative Outlier Cluster Factor(ROCF)

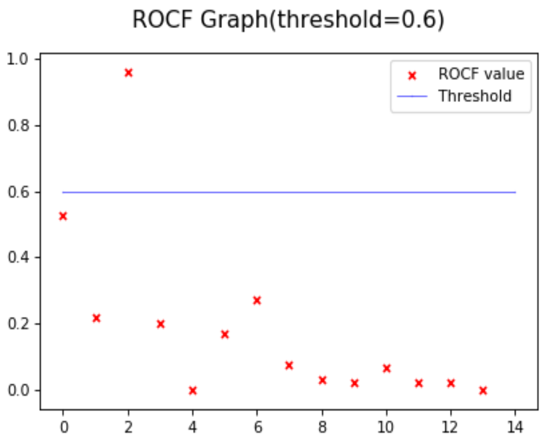
- Blobs dataset(n_samples=250, centers=5, n_features=2, random_state=3)

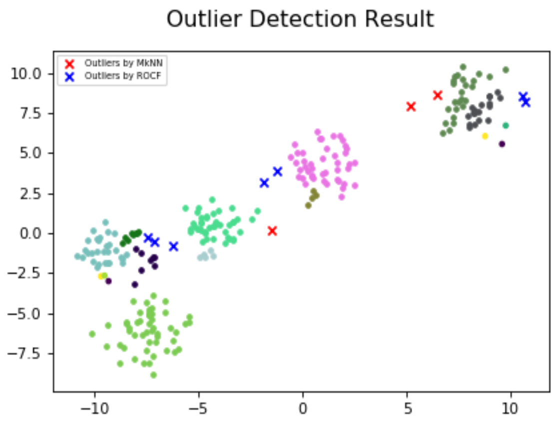


Relative Outlier Cluster Factor(ROCF)

- Blobs dataset(n_samples=250, centers=5, n_features=2, random_state=3)

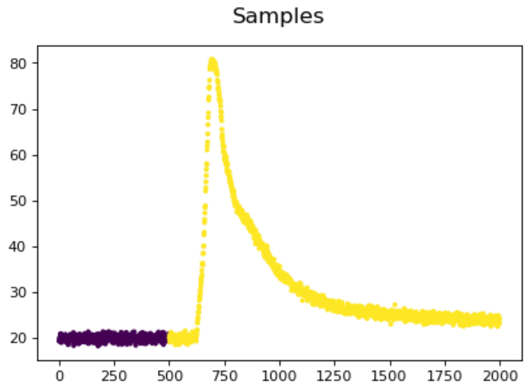






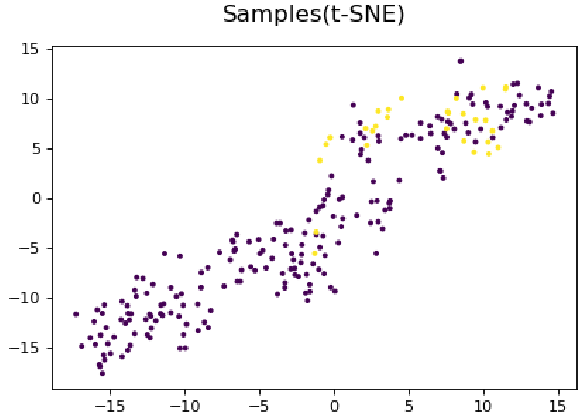
Relative Outlier Cluster Factor(ROCF)

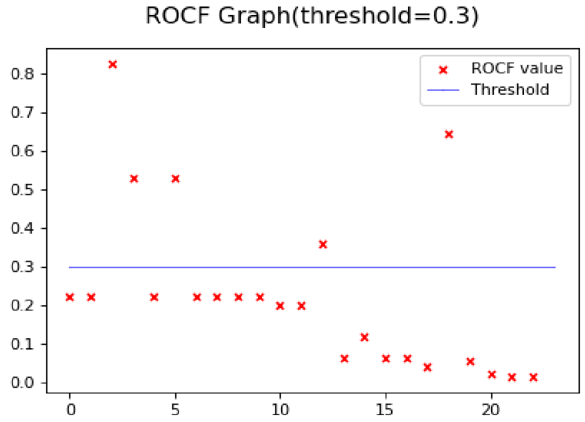
- Fire dataset
- Outliers are not detected.

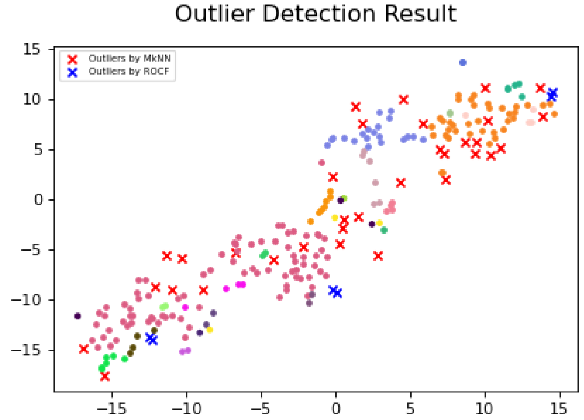


Relative Outlier Cluster Factor(ROCF)

- Vertebral dataset(240 data points, 5 features, 30 outliers)
- k= 4, ROCF threshold= 0.3
- Precision: 0.23
- Recall: 0.18

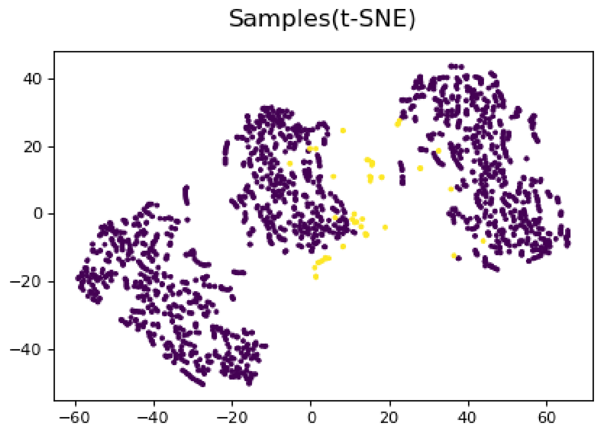


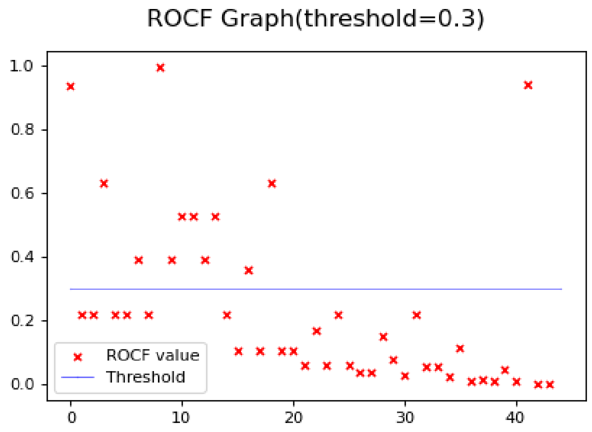


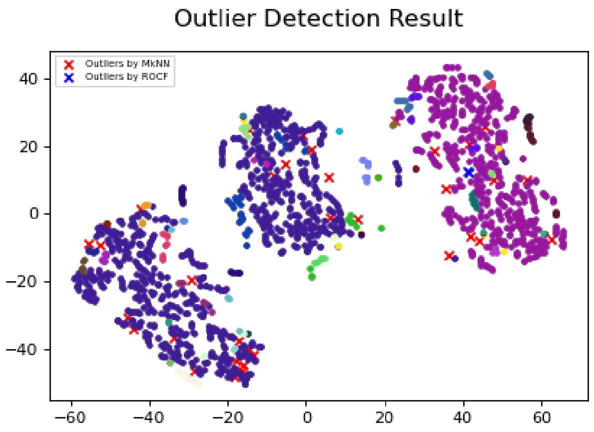


Relative Outlier Cluster Factor(ROCF)

- Vowels dataset(1456 data points, 12 features, 50 outliers)
- k= 4, ROCF threshold= 0.3
- Precision: 0.2
- Recall: 0.28







References

- Maier, Hein, Luxburg. (2009). Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. Theoretical Computer Science, Volume 410, Issue 19, pp1749-1764.
- Huang et al. (2017). A novel outlier cluster detection algorithm without top-n parameter. Knowledge-Based Systems, Vol.121. pp32-40.