# MultiKOC: Multi-One-Class Classifier Based K-Means Clustering

By (Group 1) - Ayush Sharma(2019AAPS0484G)
Rachit Jain (2019A7PS0140G)

# Acknowledgement

We would like to take this opportunity to express our deep sense of gratitude to everyone who supported us in our work during this project.

We have tried to implement the paper “MultiKOC:Multi-One-Class Classifier Based K-Means Clustering” written by Loai Abdallah, Murad Badarna, Waleed Khalifa and Malik Yousef.

We would like to thank Dr. Harikrishnan N B for giving us the opportunity to implement different Machine Learning concepts. We would also like to thank our mentor Ramanathan for his supportive nature and for providing us with any help required to complete our project.

# Introduction

**The motivation for using MultiKOC (Multi-one-class classifier based on K-means) is that multiple subclusters can occur in the class cluster of one-class classification problems which classic one-class classifiers fail to handle as they do not see the negative samples or sub-data, especially in computational biology such as multiple tumors, protein folds, biometrics, MRI, etc. The MultiKOC handles this issue by clustering the positive samples using K means and then training a one-class classifier for each cluster.**

# Problem Definition

- In many real world problems with two-class data, one of the clusters consists of multiple subclusters.
- Examples include but are not limited to, breast cancer data, thyroid gland data, iris classification dataset, etc.
- But, simple one-class classification assumes that the data only consists of two pure compact clusters.
- This leads to insufficient and less accurate results.
- Hence, a more reliable method for classification is needed.

# Objective

To test whether MultiKOC can use the hidden information of the dataset to produce better results than classic one class and two class classifiers.

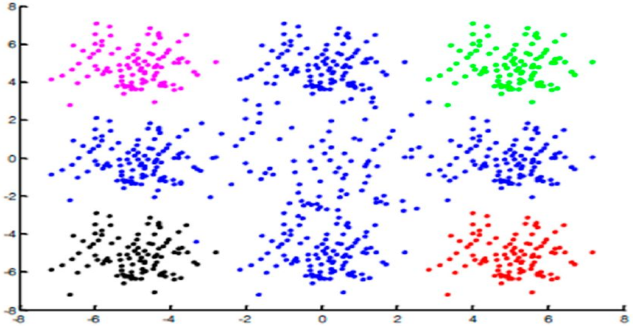


Figure :- The positive class consists of four subgroups. The negative class is in blue color. Each cluster has a different color (pink, green, black, and red).

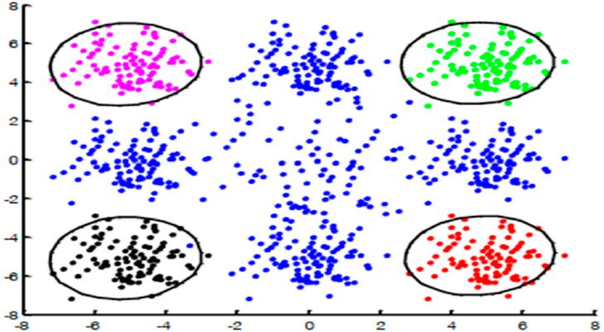


Figure :- The MultiKOC trained over the positive examples. As can be seen, the positive examples are classified into four different clusters.

# Methodology

The MultiKOC uses a simple methodology. It first takes into account only the positive samples, disregarding the negative samples. Then it uses K-means clustering (other clustering algorithms can be used according to the dataset) to divide the positive samples into clusters. The number of clusters is not critical, as identifying 2 different clusters as 1 is more problematic than dividing one cluster into 2 clusters. After clustering, one-class classifiers (J48, Random Forest, Naive Bayes, SVM, etc.) are trained on each cluster.

Now, for a test instance, all the one-class classifiers are used, and if at least one of them identifies it as positive, then it is considered as positive, else negative.

# MultiKOC Classifier Algorithm.

1. Select k - the number of subsets.
2. Apply the K-means clustering algorithm over the positive class (apply on the examples of the training set);
3. For each cluster build a one class classifier.
4. Given an unlabeled instance x;
5. Let class ← negative;
6. For each classifier clfi do;
    a. If clfi (x) is positive then
        i. class ← positive

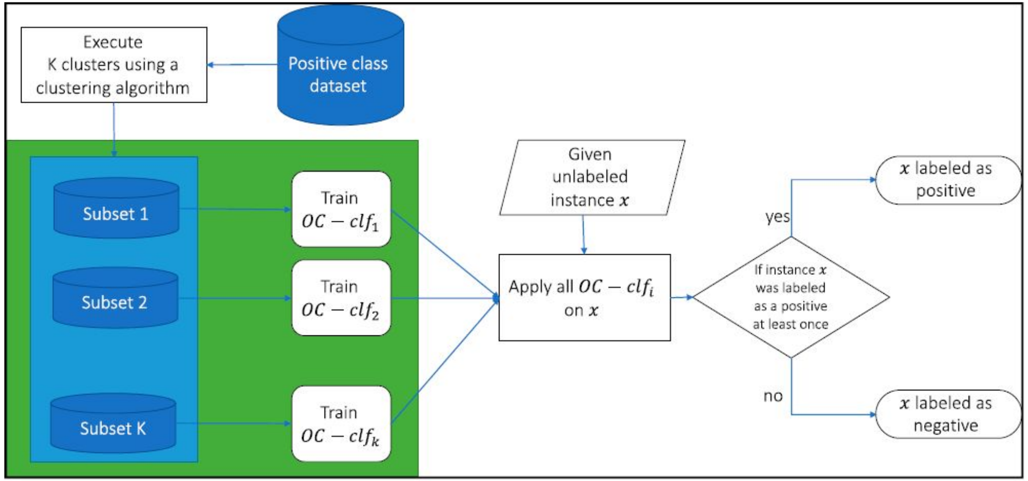
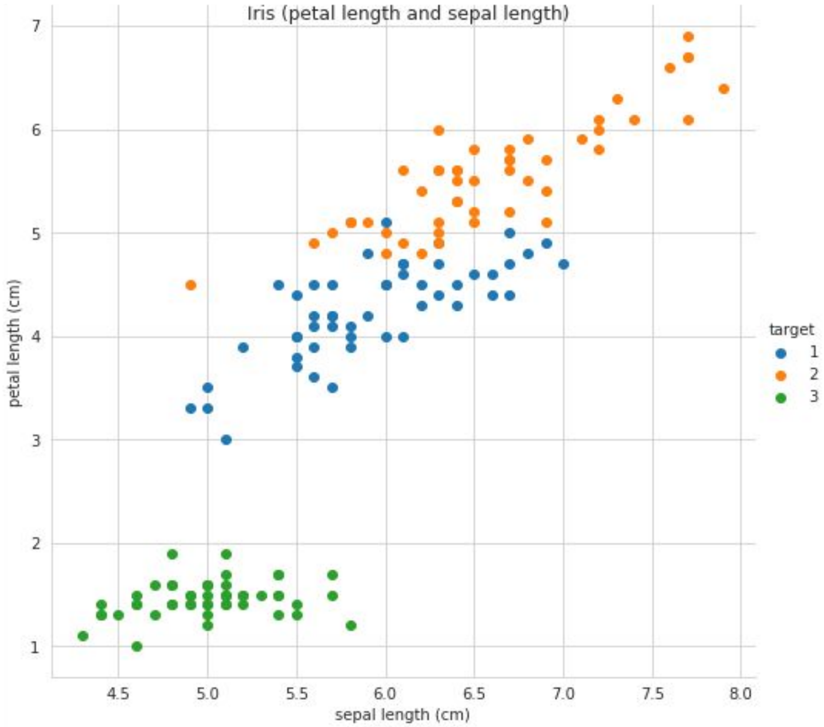


Fig: Flow diagram of MultiKOC Algorithm

# Datasets Used

*In the paper, 3 datasets were used - Iris, New Thyroid, and Syntactic. Syntactic dataset was not available, so we have used Wisconsin Breast Cancer dataset instead. Also we have derived an extra dataset from Iris.

# Iris Dataset



This data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length and width, stored in a 150x4 numpy.ndarray. It contains 50 data points of each species. Dataset is present in sklearn's toy dataset.

**Attributes: petal length (cm), petal width (cm), sepal length (cm), sepal width (cm)**

**Fig. Plotted Iris dataset w.r.t. 2 features - petal length and sepal length.**
**class 1 - versicolor**
**class 2 - virginica**
**class 3 - setosa**
**1 x-unit = 0.5 cm**
**1 y-unit = 1 cm**

# Binary datasets derived from Iris

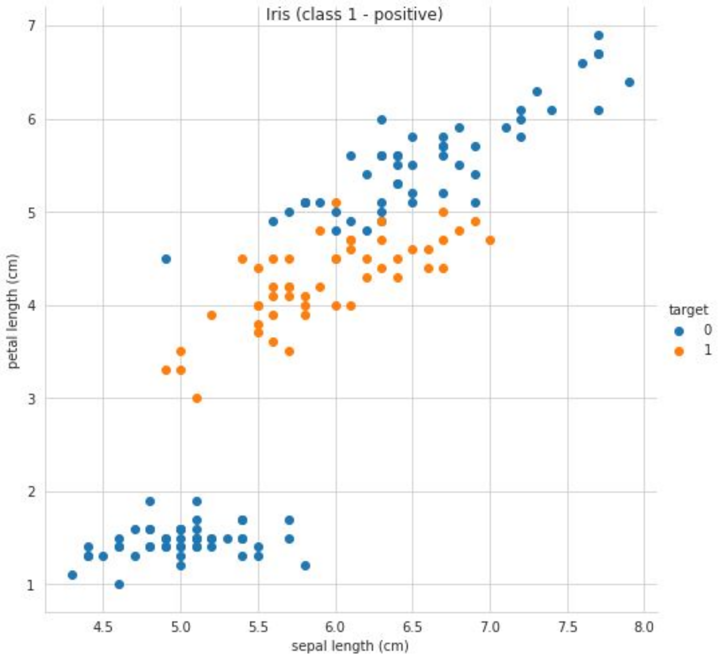


Fig. Iris dataset with class 1 positive(50 points), 2,3 negative(100 points)
1 x-unit = 0.5 cm
1 y-unit = 1 cm

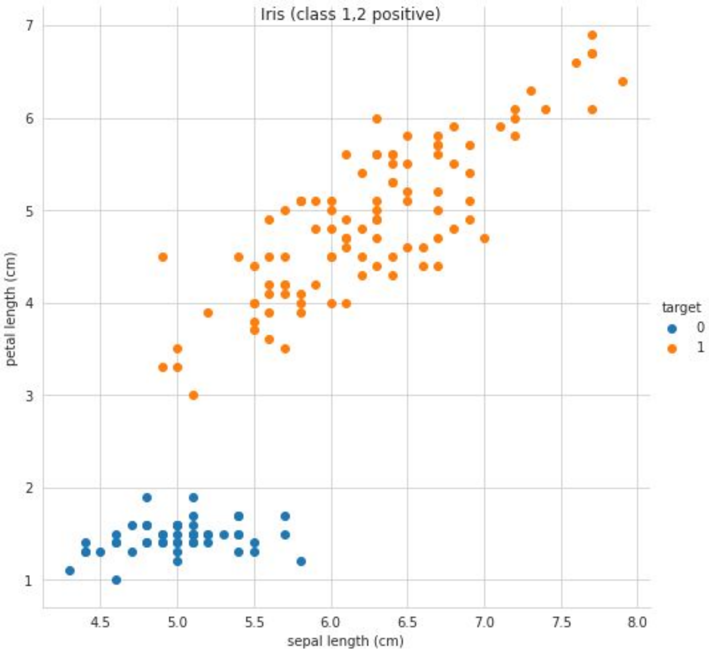


Fig. Iris dataset with class 1,2 positive(100 points) ,3 negative(50 points)
1 x-unit = 0.5 cm
1 y-unit = 1 cm

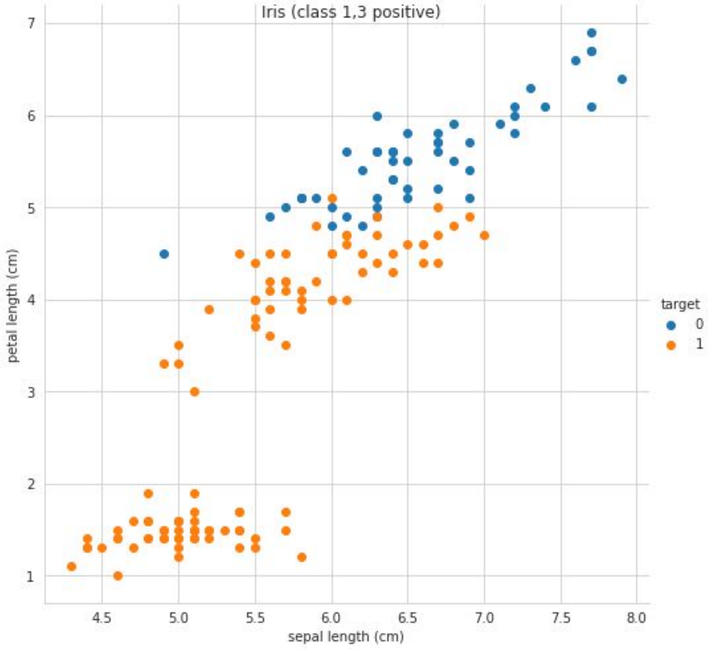


Fig. Iris dataset with class 1,3 positive(100 points), 2 negative(50 points)
1 x-unit = 0.5 cm
1 y-unit = 1 cm

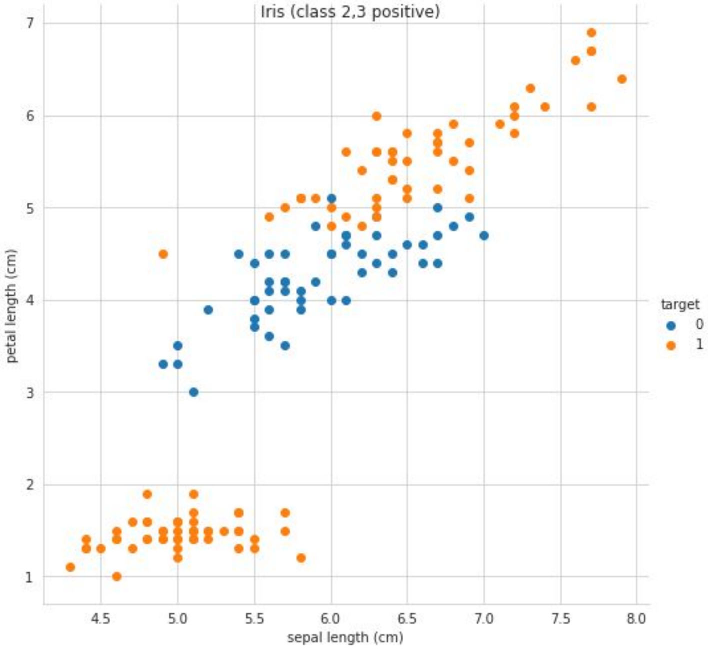
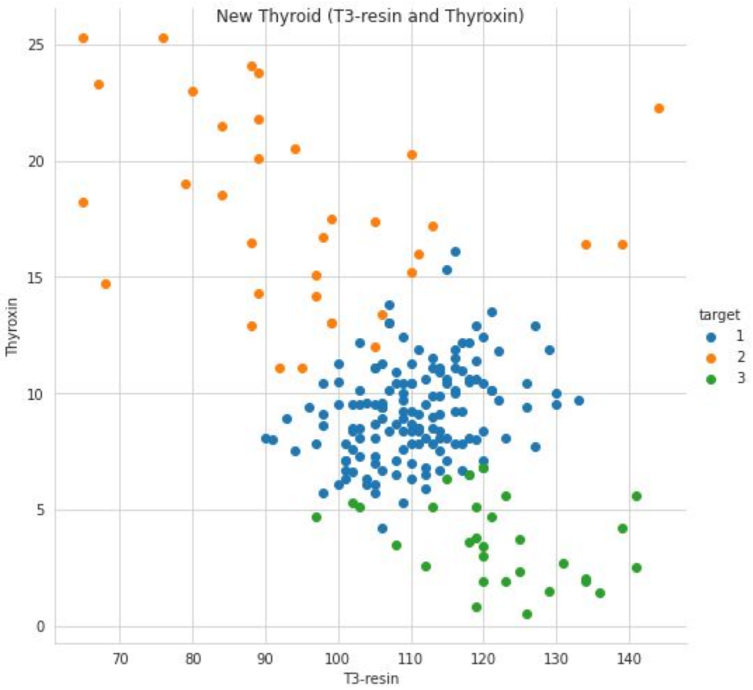


Fig. Iris dataset with class 2,3 positive(100 points), 1 negative(50 points)
1 x-unit = 0.5 cm
1 y-unit = 1 cm

# New Thyroid Dataset



Five lab. tests are used to try to predict whether a patient's thyroid to the class euthyroidism, hypothyroidism or hyperthyroidism. The diagnosis (the class label) was based on a complete medical record, including anamnesis, scan etc. Contains 150 normal, 35 hyper, 30 hypo data points.

**Attributes: T-3 resin - T3-resin uptake test. (A percentage), thyroxin - Total Serum thyroxin as measured by the isotopic displacement method, triiodothyronin - Total serum triiodothyronine as measured by radioimmuno assay, TSH - basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay, TSH diff - Maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.**

**Fig. Plotted New Thyroid dataset w.r.t. 2 features - T3-resin and Thyroxin.**
**class 1 - normal**
**class 2 - hyper**
**class 3 - hypo**
**1 x-unit = 10 % T3 resin uptake**
**1 y-unit = 5 µg/dL of blood**

# Binary datasets derived from New Thyroid

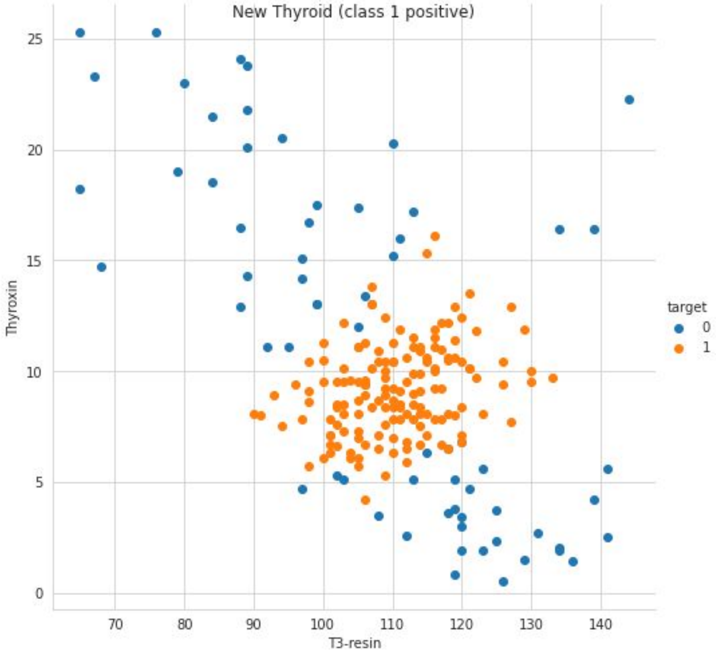


**Fig. New thyroid dataset with class 1 positive(150 points), 2,3 negative (65 points)**
**1 x-unit = 10 % T3 resin uptake**
**1 y-unit = 5 µg/dL of blood**

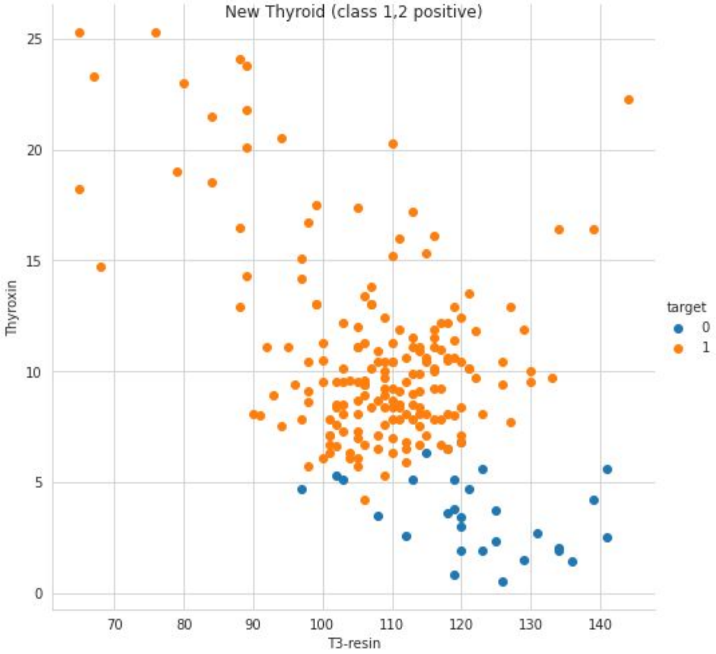


**Fig. New thyroid dataset with class 1,2 positive (185 points), 3 negative (30 points)**
**1 x-unit = 10 % T3 resin uptake**
**1 y-unit = 5 µg/dL of blood**

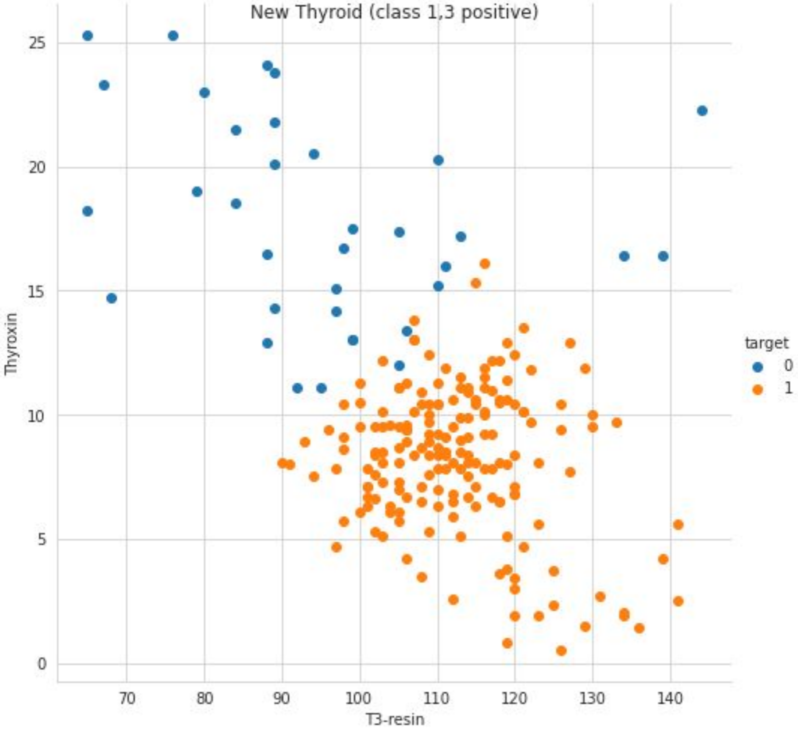


Fig. New thyroid dataset with class 1,3 positive (180 points),
2 negative (35 points)
1 x-unit = 10 % T3 resin uptake
1 y-unit = 5 µg/dL of blood

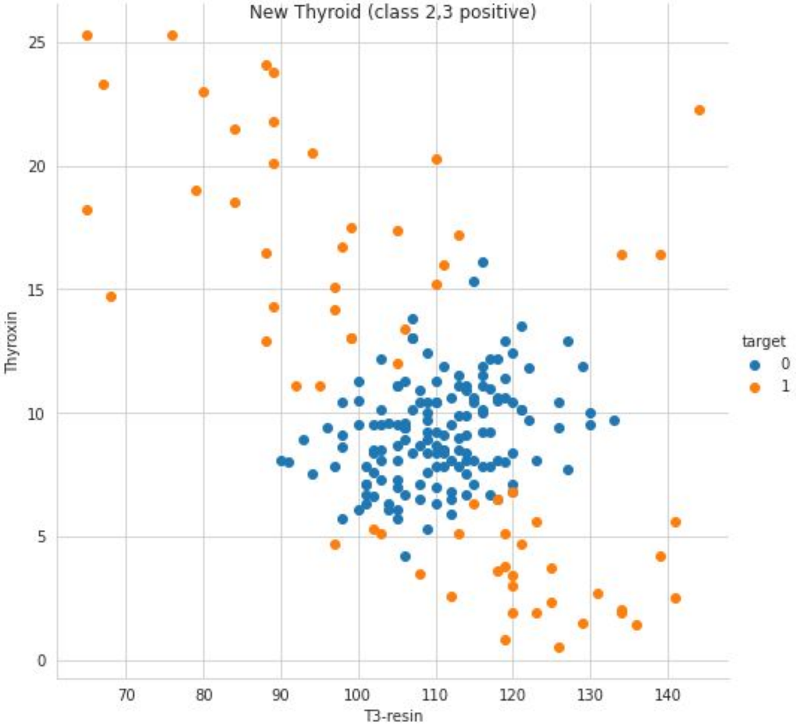
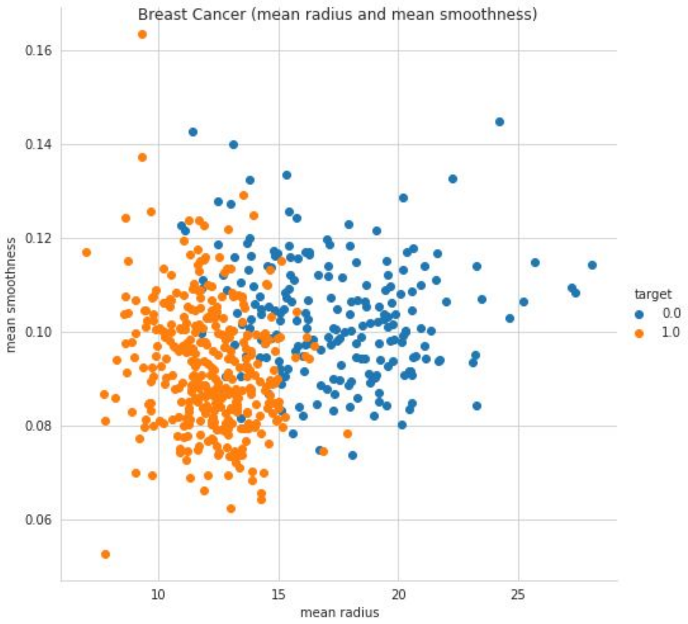


Fig. New thyroid dataset with class 1,2 positive (150 points),
3 negative (65 points)
1 x-unit = 10 % T3 resin uptake
1 y-unit = 5 µg/dL of blood

# Wisconsin Breast Cancer Dataset



This is a copy of UCI ML Breast Cancer Wisconsin (Diagnostic) datasets. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Contains 569x30 dataset with 2 classes - Benign (357 data points) and Malignant (212 data points).

**Attributes: radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1) (mean, best, and worst of each)**

**Fig. Plotted Wisconsin Breast Cancer dataset w.r.t. 2 features - mean smoothness and mean radius.**
**class 1 - Benign**
**class 2 - Malignant**
**1 x-unit = 5 um**
**1 y-unit = 0.02  um**

# Experiments and Results

# Parameters And Other Details

- In the paper, 4 one class classifiers were trained on 80% of positive data, and tested on 20%positive data+negative data.(Shown as Positive in table below) However, the working/ models of the one class classifiers are not mentioned.
- We have tried the following classifiers in our code:
  - One class SVM - sklearn's OneClassSVM module
  - One class Naive Bayes - sklearn's ComplementNB and GaussianNB modules
  - One class Random Forest - sklearn's IsolationForest module
  - One class J48 - sklearn's DecisionTreeClassifier module
- Since these may not be the exact models, the results slightly vary from paper's results.
- We have also experimented with imbalanced datasets, i.e. using some of the negative class samples in training to see differences in results. (Shown as Positive+Negative in table below)
- We have used Accuracy, Precision, Recall, and Macro F1 score as metrics. F1 score is displayed in next slide, rest can be seen in our attached python notebook.

|  | Clusters | J48<br>(F1 Score) |  | SVM<br>(F1 Score) |  | Random Forest<br>(F1 Score) |  | Naive Bayes<br>(F1 Score) |  |
|---|---|---|---|---|---|---|---|---|---|
| Training Data |  | Positive + Negative | Positive | Positive + Negative | Positive | Positive + Negative | Positive | Positive + Negative | Positive |
| class_1_iris | 3 | 0.083 | 0.083 | 0.471 | 0.551 | 0.316 | 0.097 | 0.083 | 0.083 |
| class_1_2_iris | 3 | 0.222 | 0.222 | 0.471 | 0.611 | 0.457 | 0.231 | 0.222 | 0.222 |
| class_2_3_iris | 3 | 0.222 | 0.222 | 0.420 | 0.636 | 0.478 | 0.339 | 0.222 | 0.222 |
| class_1_3_iris | 3 | 0.222 | 0.222 | 0.413 | 0.641 | 0.222 | 0.234 | 0.222 | 0.222 |
| class_1_thyroid | 2 | 0.240 | 0.240 | 0.427 | 0.480 | 0.547 | 0.246 | 0.240 | 0.240 |
| class_1_2_thyroid | 2 | 0.356 | 0.356 | 0.373 | 0.583 | 0.395 | 0.361 | 0.356 | 0.356 |
| class_2_3_thyroid | 5 | 0.074 | 0.074 | 0.480 | 0.644 | 0.165 | 0.105 | 0.074 | 0.074 |
| class_1_3_thyroid | 2 | 0.336 | 0.336 | 0.387 | 0.682 | 0.367 | 0.468 | 0.336 | 0.336 |
| breast_cancer_df | 3 | 0.207 | 0.202 | 0.432 | 0.456 | 0.252 | 0.206 | 0.202 | 0.202 |

# Conclusion And Discussions

- The current results show that it is possible to build up a multi-one-class classifier with a combined clustering beforehand process based only on positive examples yielding a significant improvement over the one-class and similar results as the two-class.
- The MultiKOC approach based on the one-class classification method discussed in this paper is based on partitioning the training data into different clusters and constructing the one-class model for each cluster.
- The MultiKOC approach performs similarly to that of the two-class version but it includes more interpretable classifiers which help in performing deep analysis and exploring the hidden structure of the data.

# Future Works

- The current framework is tested only for classification type data. It could also be tested for other data types.
- Second, the concept used in this approach can also be applied to other types of classifiers.
- In this framework, after checking with all the classifiers, we assign positive or negative class to the data point. Instead, a voting mechanism based on the results of multi one-class classifiers can be devised which assigns the label to this data point.

# Challenges

- The nature of the metrics used to produce the results were not mentioned in the paper.
- It was mentioned that the tests were run 100 times and then average was taken to report the results but what parameters were changed in those 100 times was not mentioned.
- The results from different types of one classifiers was also mentioned in the paper but which One Class classifiers were used and what parameters were used for those classifiers was not mentioned.
- There were no links to dataset used in the paper.

# Code

Github Link

# References

- Paper → MultiKOC:Multi-One-Class Classifier Based K-Means Clustering
- Datasets →
  - Iris Dataset → Link
  - New Thyroid Dataset → Link
  - Wisconsin Breast Cancer Dataset → Link

# Thank You!!