INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES MONTERREY

CS5051. COMPUTATIONAL TECHNIQUES IN MACHINE LEARNING

ASSIGNMENT 5: ANOMALY DETECTION PROBLEMS

PROFESSOR: MIGUEL ANGEL MEDINA PÉREZ

JORGE FRANCISCO CIPRIÁN SÁNCHEZ A01373326
ERICK ALAN BARLANDAS QUINTANA A01129279

NOVEMBER 3RD, 2021.

The student evaluates and analyses the results of the algorithms BRM, GMM, and ISOF according to AUC, without scaling or normalizing the database, using MinMax scaling, and using Standard normalizing, in at least 60 of the databases provided by the professor. The student shows the box plot, the Friedman test with a posthoc test, visualizes a CD diagram with the results of a statistical test, and discusses the obtained results of the evaluated algorithms. The student modifies BRM implementation, so it can work with any dissimilarity measure and tries BRM with at least three dissimilarity measures

For this assignment, we were required to evaluate and analyze the results of the algorithms BRM, GMM and ISOF:

- Including the MinMax Scaling for each algorithm
- Including the Standard normalizing for each algorithm
- Modify the BRM implementation, so it can work with any dissimilarity measure.

At the end, we tested the following algorithms:

1. ISOF
2. ISOF_stand
3. ISOF_minmax
4. GMM
5. GMM_stand
6. GMM_minmax
7. BRM_euc
8. BRM_euc_stand
9. BRM_euc_minmax
10. BRM_cos
11. BRM_cos_stand
12. BRM_cos_minmax
13. BRM_man
14. BRM_man_stand
15. BRM_man_minmax
16. ocSVM

where:
stand = standard normalized
minmax = MinMax scaling
euc = Euclidean distance
man = Manhattan distance
cos = Cosine distance

We tested the previous algorithms according to AUC, in 60 different databases, obtaining the boxplot, the Friedman test with a posthoc test (Nemenyi test), and created the CD diagram with the results of the statistical tests.

Boxplot results:

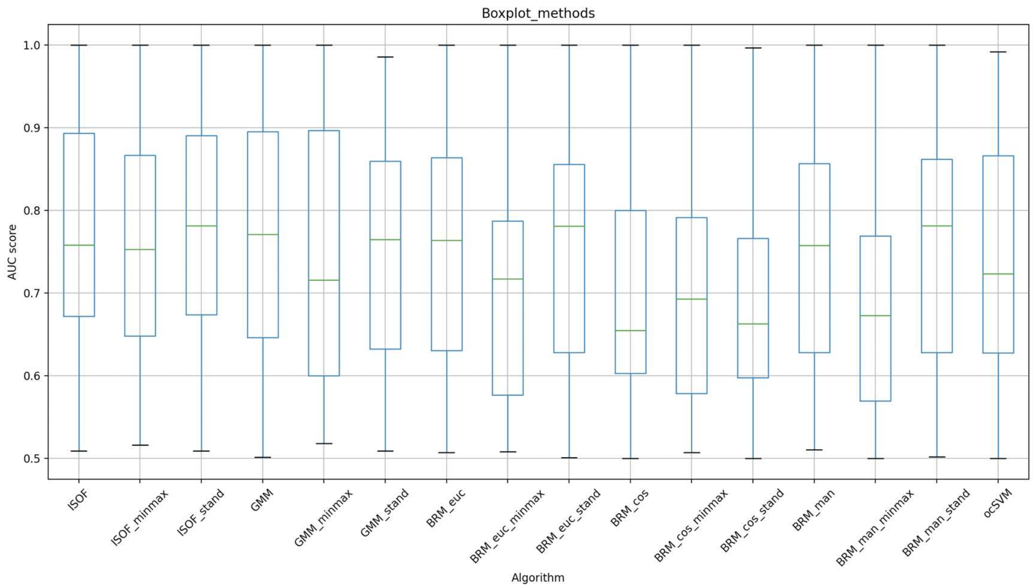

Image 1. Boxplot for the different algorithms.

After testing the different algorithms with the 60 datasets, we created a boxplot graph to compare the AUC values obtained. As it can be observed, most of the algorithms obtained a max value of 1 or close to 1, except for the GMM_stand and the ocSVM. Besides this, most algorithms got a minimum value of 0.5 or close to 0.5, except for GMM_minmax and BRM_man. The ISOF algorithm had a similar median value in its three models (around 0.75 to 0.78). The model where the median changed more was the BRM_man algorithm, with median values from 0.78 to 0.67 approximately. The algorithm with the lowest median was the BRM_cos algorithm, with a median value of 0.65, and the algorithms with the highest median value were the BRM_euc_stand, the ISOF_stand, and the BRM_man_stand algorithms, with median values of 0.78-0.79. The algorithm with the more dispersed data was the GMM_minmax algorithm, and the algorithm with the more robust data was the BRM_cos_stand algorithm. However, this algorithm has a low median value compared with the other algorithms (0.66). Summarizing, in most of the cases the standardized algorithms obtained the highest median value, except for the BRM algorithm with cosine distance as its dissimilarity metric.

Friedman test and Nemenyi test.

To perform the Friedman and posthoc test, we used the python package autorank. Autorank is a simple Python package with one task: simplify the comparison between (multiple) paired populations. Autorank determines if all populations are normal to select appropriate statistical markers for reporting. In our case, as there are more than two populations and at least one population is not normal, or the populations are heteroscedastic, we use Friedman's test with the Nemenyi post-hoc test [1]. The Nemenyi post-hoc test intended to find the groups of data that differ after a global statistical test has rejected the null hypothesis that the performance of the comparisons on the groups of data is similar. The test makes pair-wise tests of performance.

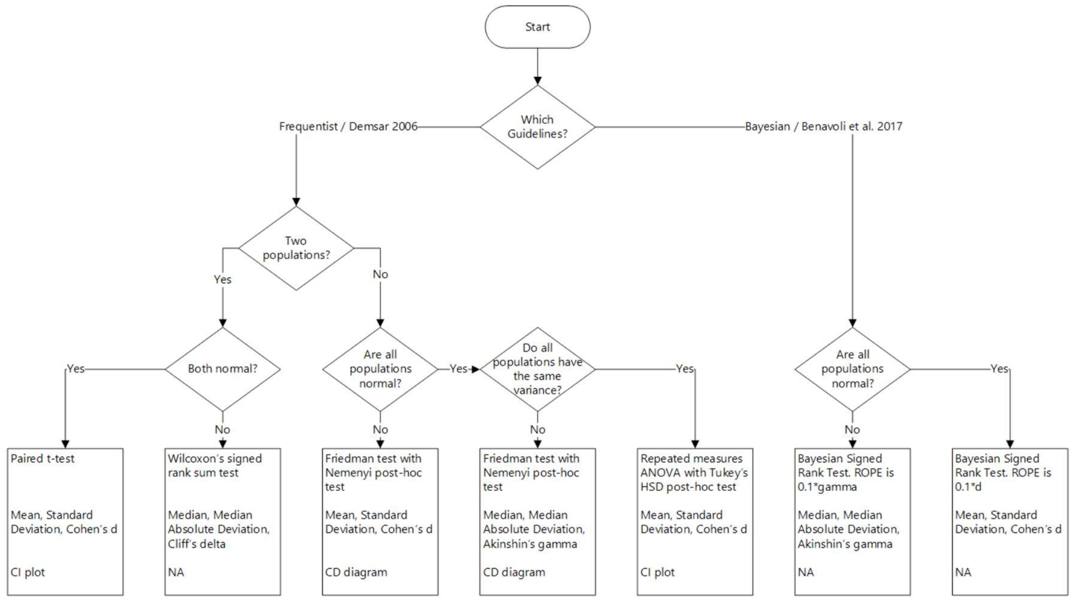

Image 2. Autorank decision flow chart. Image recovered from: https://github.com/sherbold/autorank. This image was taken without authorization.

The appropriate statistical test for the frequentist approach is determined based on the normality of the data, the homogeneity of the populations (equal variances), and the number of populations. The best test for homogeneity depends on the normality of the data. As there are more than two populations and at least one population is not normal, or the populations are heteroscedastic, we used Friedman's test with the Nemenyi post-hoc test [1].

Using autorank, the non-parametric Friedman test was carried as omnibus to determine if there were any significant differences between the median values of the populations. After that, we employ the post-hoc Nemenyi test to infer which differences were significant. The following image (image 3) contains the mean rank, the median, the median absolute deviation, the confidence intervals, and the effect size among all populations over the samples. In our case, if the difference between mean rank is greater than the critical distance (CD=2.98 of the Nemenyi test), the differences between populations are significant.

```
RankResult(rankdf=
                 meanrank    median       mad  ci_lower  ci_upper effect_size   magnitude
ISOF_stand       6.575000  0.781766  0.160377  0.673077     0.915         0.0  negligible
ISOF             7.141667  0.757991  0.166647  0.662162  0.897727    0.145374  negligible
GMM              7.708333  0.771131  0.188503  0.639344  0.941713    0.060771  negligible
GMM_stand        7.725000  0.764908  0.164821  0.627397  0.871622    0.103672  negligible
GMM_minmax       7.808333  0.715979  0.201809  0.594086  0.900896    0.360923       small
BRM_man          7.816667  0.757703  0.181896  0.622596     0.875    0.140328  negligible
BRM_man_stand    7.875000  0.781540   0.17413  0.628082  0.878049    0.001349  negligible
ISOF_minmax      7.991667  0.752841  0.162684  0.645833  0.883212    0.179065  negligible
ocSVM            8.058333  0.723648  0.185093  0.625801      0.88    0.335603       small
BRM_euc_stand    8.133333  0.781252  0.163727  0.624849     0.875    0.003174  negligible
BRM_euc          8.175000  0.763935  0.173846   0.63011  0.878788    0.106616  negligible
BRM_euc_minmax   9.400000  0.717395  0.171024  0.571429  0.846847    0.388278       small
BRM_cos_stand   10.191667  0.663209  0.126499  0.595192     0.775    0.820832       large
BRM_man_minmax  10.391667  0.673065   0.14729  0.559659  0.794788    0.705975      medium
BRM_cos         10.408333  0.654965  0.132522  0.599251  0.832011    0.861947       large
BRM_cos_minmax  10.600000  0.693059  0.159699  0.562329  0.816667    0.554287      medium
```

Image 3. Autorank results using the non-parametric Friedman test as omnibus, and the post-hoc Nemenyi test.

Table 1. Autorank attributes for the Friedman test and the post-hoc Nemenyi test.

Attribute	***Value***
p-Value	4.633416283751652e-08
Omnibus	Friedman
Posthoc	Nemenyi
All_normal	False
Homoscedastic	True
Pval_homogeneity	0.49248288028115406
Homogeneity_test	levene
cd	2.9781494590295132
Alpha	0.05
Num_samples	60

CD Diagram

The CD diagram is a great graph to compare the outcomes of multiple treatments over multiple observations. In machine learning, CD diagrams are used to compare the performance of multiple methods over multiple data sets [2]. After performing the Friedman and post-hoc Nemenyi test, we created the following CD diagram:

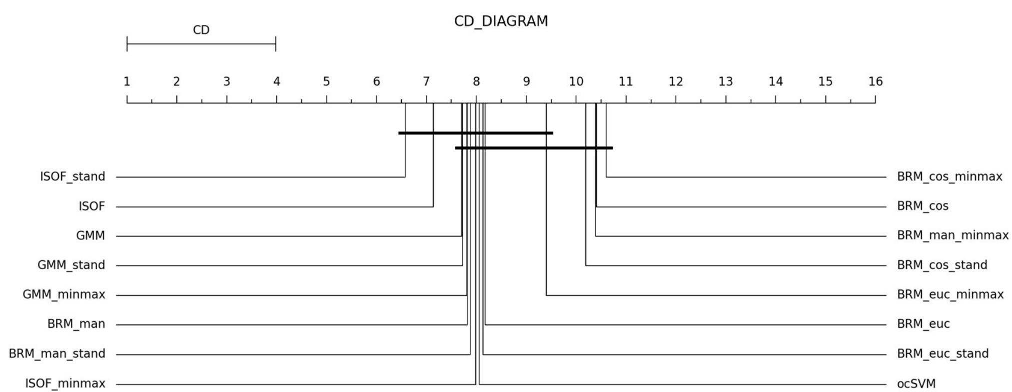

Image 4. CD diagram obtained after performing the Friedman test and the post-hoc Nemenyi test.

The name of each algorithm points to its mean rank, and the black lines perpendicular to the algorithms' lines indicate if the data is different or not from the other algorithms. For example, ISOF_stand, ISOF, GMM, GMM_stand, GMM_minmax, BRM_man, BRM_man_stand, ISOF_minmax, ocSVM, BRM_euc_stand, BRM_euc, and BRM_euc_minmax are joint by the first horizontal black line, indicating that we cannot tell that the data of these algorithms are different from each other. Nevertheless, we can say that the data from BRM_cos_minmax, BRM_cos, BRM_man_minmax, and BRM_cos_stand are different from this first group data. For the second black line perpendicular to the algorithms' lines, we can observe that only ISOF_stand and ISOF data are different from the other algorithms. The p-value obtained for this project was 4.63e-08, so we can conclude that the data is statistically significant.

References

1. Herbold, S. (2020). Autorank: A Python package for automated ranking of classifiers. *Journal of Open Source Software*, *5*(48), 2173. https://doi.org/10.21105/joss.02173
2. Github. (s. f.). *Home · CriticalDifferenceDiagrams.jl*. CriticalDifferenceDiagrams.Jl. Recuperado 1 de noviembre de 2021, de https://mirkobunse.github.io/CriticalDifferenceDiagrams.jl/dev/