A01280418 - Luis Javier Ramírez

Assignment 5 Report
One-Class Anomaly Detection in 60 databases

I. Databases

For this assignment, we decided to use the first 60 of the 95 databases that were shared. The databases used were the following:

abalone-17_vs_7-8-9-10	ecoli-0-3-4_vs_5	glass2
abalone19	ecoli-0-3-4-6_vs_5	glass4
abalone-19_vs_10-11-12-13	ecoli-0-3-4-7_vs_5-6	glass5
abalone-20_vs_8-9-10	ecoli-0-4-6_vs_5	glass6
abalone-21_vs_8	ecoli-0-6-7_vs_3-5	habermanImb
abalone-3_vs_11	ecoli-0-6-7_vs_5	iris0
abalone9-18	ecoli1	kr-vs-k-one_vs_fifteen
car-good	ecoli2	kr-vs-k-three_vs_eleven
car-vgood	ecoli3	kr-vs-k-zero_vs_eight
cleveland-0_vs_4	ecoli4	kr-vs-k-zero_vs_fifteen
dermatology-6	flare-F	kr-vs-k-zero-one_vs_draw
ecoli-0_vs_1	glass0	led7digit-0-2-4-5-6-7-8-9_vs_1
ecoli-0-1_vs_2-3-5	glass-0-1-2-3_vs_4-5-6	lymphography-normal-fibrosis
ecoli-0-1_vs_5	glass-0-1-4-6_vs_2	new-thyroid1
ecoli-0-1-3-7_vs_2-6	glass-0-1-5_vs_2	new-thyroid2
ecoli-0-1-4-6_vs_5	glass-0-1-6_vs_2	page-blocks0
ecoli-0-1-4-7_vs_2-3-5-6	glass-0-1-6_vs_5	page-blocks-1-3_vs_4
ecoli-0-1-4-7_vs_5-6	glass-0-4_vs_5	pimaImb
ecoli-0-2-3-4_vs_5	glass-0-6_vs_5	poker-8_vs_6
ecoli-0-2-6-7_vs_3-5	glass1	poker-8-9_vs_5

II. Algorithms

We used 4 one-class anomaly detection algorithms in the 60 databases. This were Bagging-Random Miner (BRM), Gaussian Mixture Model (GMM), Isolation Forest (ISOF) and One-Class Support Vector Machine (ocSVM). All algorithms were run with their default parameters, except for BRM, which we modified to use different dissimilarity measures, which will be explained further in Section IV.

In total, we ran 6 algorithms:

- GMM
- ISOF
- ocSVM
- BRM (Euclidean)
- BRM (Manhattan)
- BRM (Cosine)

The approximate running time of the 6 algorithms in the 60 databases was around 20-30 minutes, and we used the ROC AUC metric to evaluate them.

III. Data Transformation

Most models cannot work with categorical data, so we performed One Hot Encoding in all databases with categorical data in it to avoid this problem.

After that, we performed MinMax scaling to all the variables, as well as Standard Normalizing. This means that, for the 6 algorithms that were tested, we tested each 3 times for each database, with the different data transformations. We had 18 models tested (ex. BRM (Euclidean), BRM (Euclidean)+MinMax scaling, and BRM (Euclidean)+Standard Scaling, for each database).

Ultimately, we have 6 algorithms, with 3 different data transformations in 60 dataframes, so we ended up training 1,080 different one-class anomaly detection models. We can see the comparison of the average AUC for each of the 18 models in Figure 2.

IV. Dissimilarity Measure

One of the algorithms we tested is Bagging-Random Miner (BRM), which in its source code only uses Euclidean distance as a dissimilarity measure when training the model. We modified the BRM implementation so we could specify which dissimilarity measure we wanted to use. We ran the algorithm using 3 different measures: Euclidean, Manhattan, and Cosine distances.

V. Statistical Tests

After performing the 1,080 models, we want to know if there is an algorithm, or a combination of algorithm + Data Transformation that clearly outperforms the others. In order to confirm this, we first performed a Friedman test to check if the distributions of our algorithms are different. We obtained a p-value of 0.0008 in the Friedman test, so we rejected the null hypothesis that distributions were equal.

After this, we performed a post hoc test to find the differences. We ran a Nemenyi test, which allows us to find groups in the data that differ to other groups. We can see a Critical Difference diagram in Figure 1, which shows us that the only real statistical difference between models appears with BRM(cosine) vs ISOF and ISOF_MinMax.

VI. Visualizations

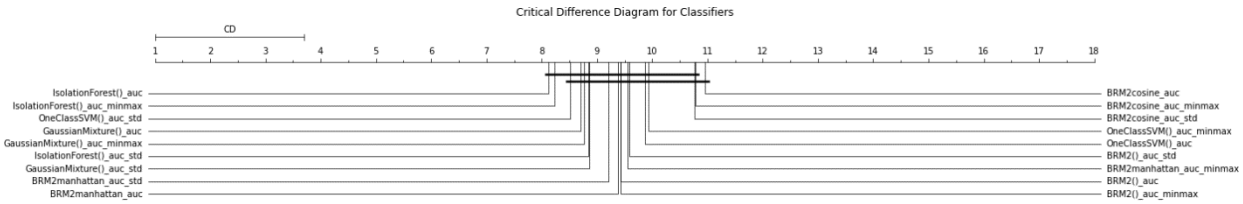

Figure 1: Critical difference diagram using Nemenyi test to compare classifiers

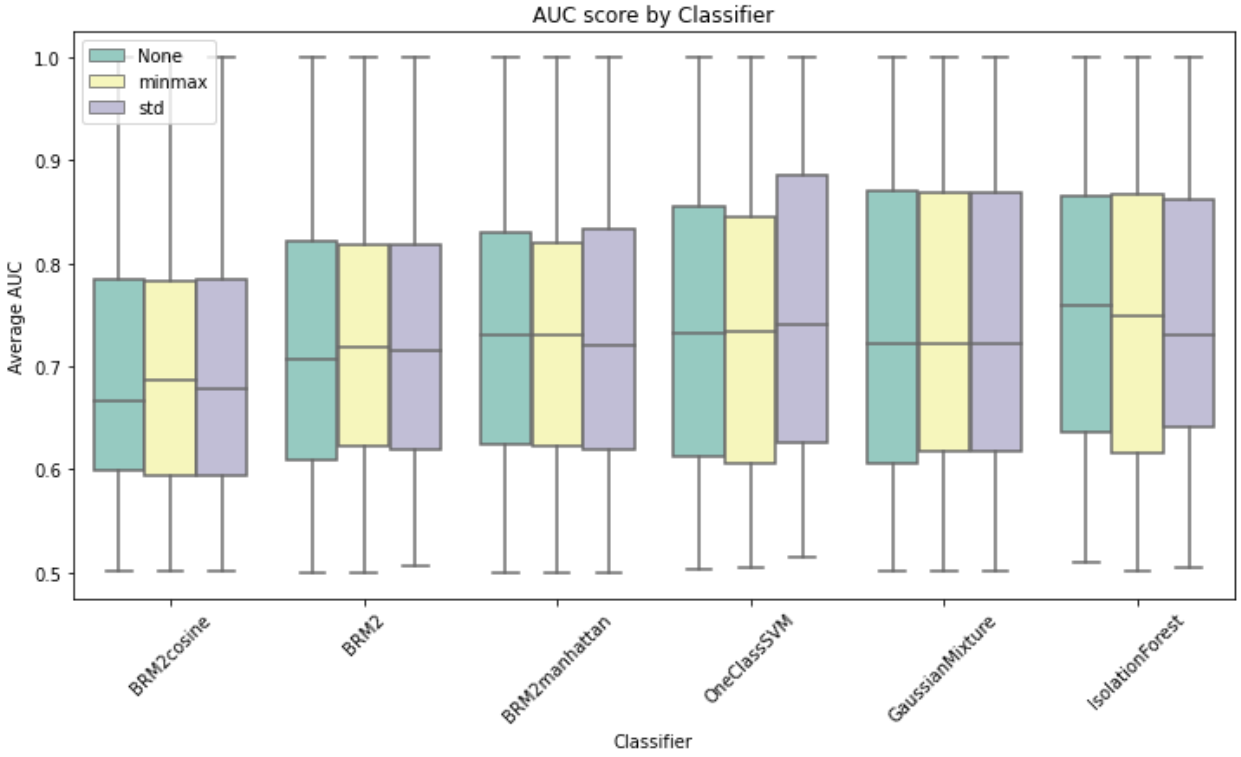

Figure 2: Box plot comparing the Average AUC between all combinations of Algorithm+Data Transformation.

VII. Discussion and conclusions

In this assignment we learned how to do a correct and robust comparison between different algorithms: testing in a lot of databases and then performing a statistical analysis.

For this assignment, after analyzing the graphs and the results from the statistical analysis, we concluded that there is not a statistical difference between BRM, ocSVM, ISOF and GMM, at least in the subset of databases that we arbitrarily selected. The only model that performed statistically worse is BRM with Cosine distance as a dissimilarity measure.

Source code for this assignment can be found in: https://github.com/luisjrmz/BRM .