A COMPARISON STUDY ON STATE-OF-THE-ART MINORITY CLASS DATA OVERSAMPLING TECHNIQUES FOR IMBALANCED LEARNING

Han Tang

TU Dublin, School of Computing

MSc in Computing, Data Analytics

January, 2020

LITERATURE REVIEW & BACKGROUND

- Imbalanced learning Approaches: Data level, algorithmic level
- Algorithmic level approaches: Cost-sensitive learning
- Data level: Under-sampling, Over-sampling, Combination of under-sampling and over-sampling
- Over-sampling: Synthetic Minority Over-sampling Technique (SMOTE)
- SMOTE still have problems, thus extensions are proposed aim to improve its performance

- Tested SMOTE Extensions families:
  - Range Restricted SMOTE extensions
    - Borderline SMOTE
    - Safe-Level SMOTE
  - Clustering Based SMOTE extensions
    - Cluster Based Synthetic Oversampling (CBSO)
    - Agglomerative Hierarchical Clustering (AHC)
    - Majority Weighted Minority Oversampling TEchnique (MWMOTE)
  - Combination of data sampling method
    - SMOTE-Edited Nearest Neighbour (ENN)
    - SMOTE-Tomek Links

GAPS & MOTIVATION

- No experiment conducted comparing different approaches on the same imbalanced dataset
- Most tests on a small number of datasets (2-3), can not give enough confidence to prove the proposed method is universally better.
- The motivation of this research is to compare those state-of-the-art SMOTE extensions and to provide a statistically significant result.

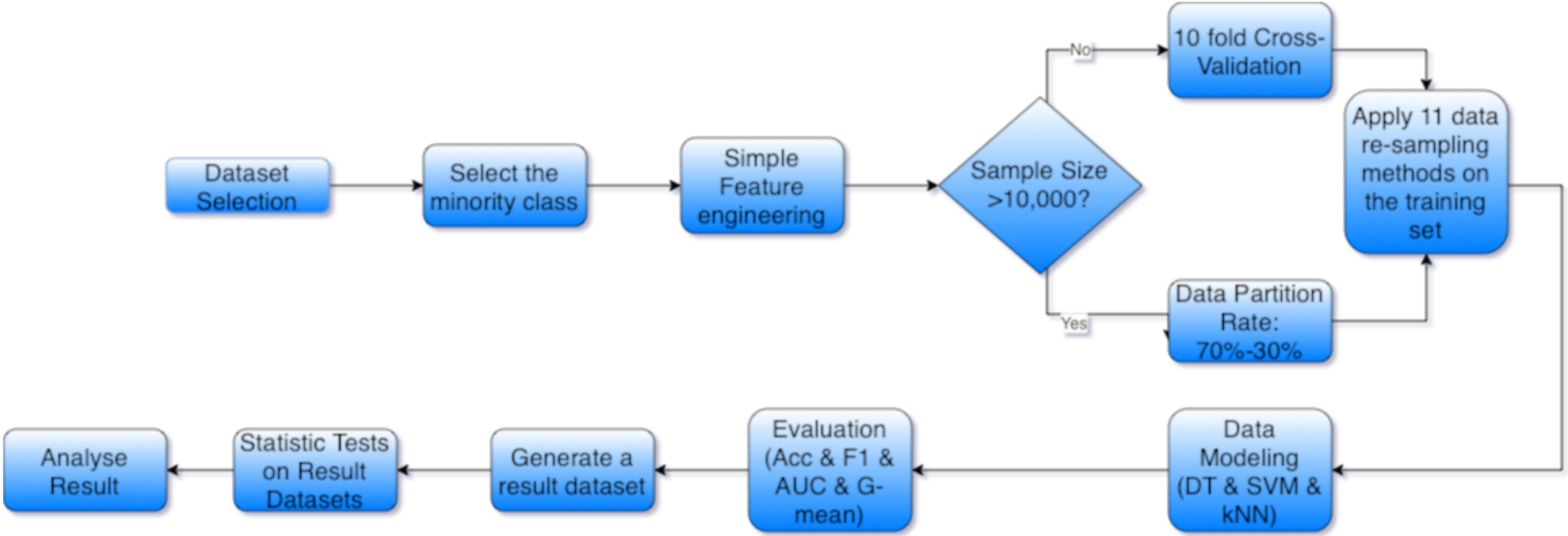
RESEARCH QUESTION

- Is there a statistically significant difference in the performances of the combination of data resampling method, range-restricted SMOTE extensions, and clustering-based SMOTE extensions when measured by F1 and AUC?

RESEARCH METHODOLOGY

"Secondary, empirical research that seeks to provide an inductive basis for future work by comparing three families of SMOTE extensions."

- Secondary : The data are gathered from existing repositories
- Empirical : The data are actual (not fabricated) and measurable
- Inductive : The result is obtained by comparing the results of the experiments.



WORKFLOW FOR EXPERIMENT

SUMMARY OF RESULTS

- The result tested by Kruskal-Wallis test and ANOVA indicates kNN and Decision Tree performs better on imbalanced learning tasks, compared to SVM.
- Within the range-restricted SMOTE extensions, though statistic test does not show a statistically significant difference, Safe level SMOTE generally produces the worst result.

SUMMARY OF RESULTS

- Within the cluster-based SMOTE extensions family, there is no method performs universally better than the others. Statistic test also indicates that their difference is not statistically significant.
- SMOTE-ENN and SMOTE-TomekLinks are not statistically significantly different.
- Different families of SMOTE extensions do not show statistically significant different results.
- When learning from unstructured imbalanced data, range-restricted SMOTE extensions could not give the best result.

CONTRIBUTION & IMPACT

- Compare the SMOTE extensions in works of literature on a large number of datasets, in a statistically significant way. Able to prove there is no any data oversampling method is the best for all cases.
- Give instructions on selecting oversampling methods for learning from unstructured data.

FUTURE WORK & RECOMMENDATIONS

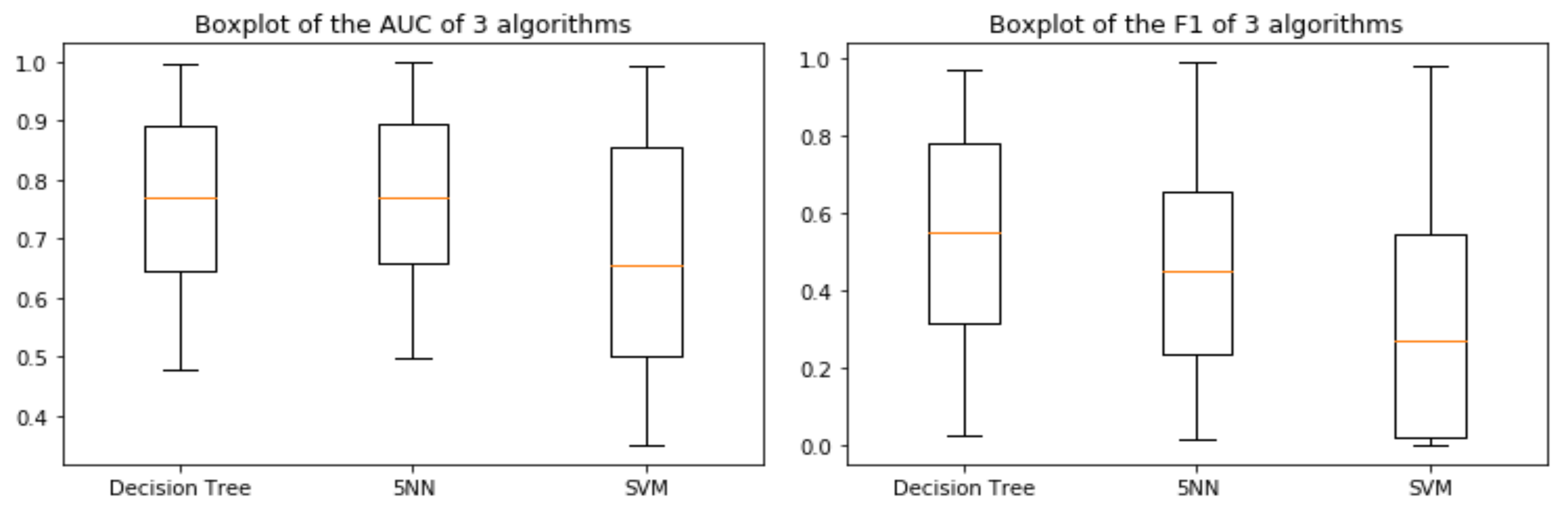
- Explore how can SMOTE be applied to deep learning tasks (high-resolution images, text)
- Explore the relations between data distributions and the performances of over-sampling methods.
- Give further recommendations to select the over-sampling methods according to data distribution patterns.

Q&A

Thank you

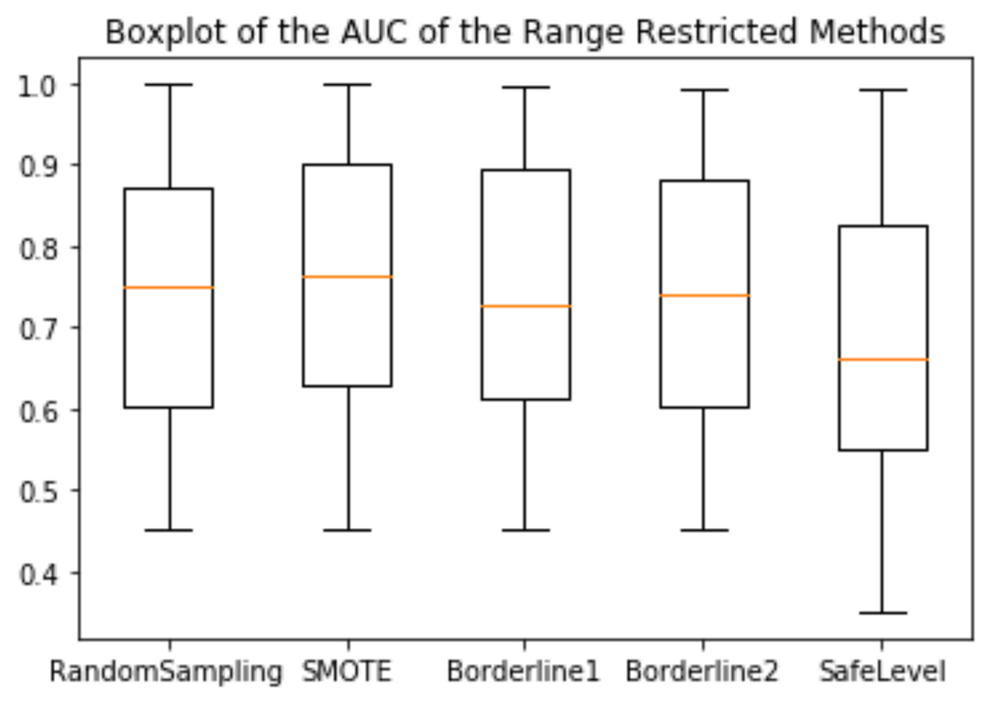
Appendix

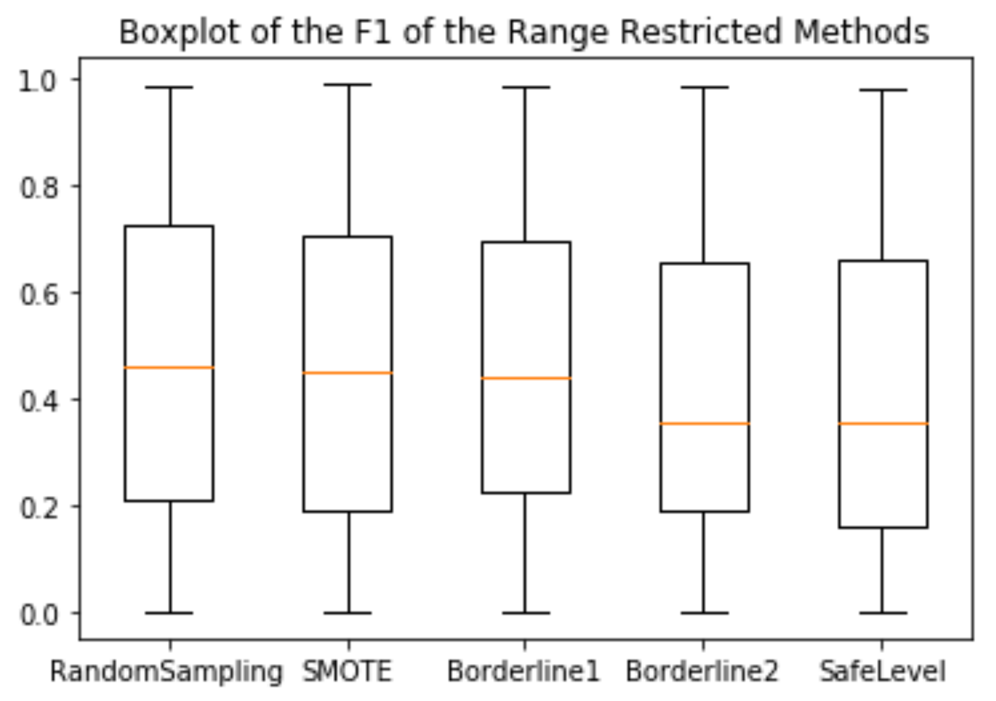
BOXPLOT OF 3 ALGORITHMS



Appendix

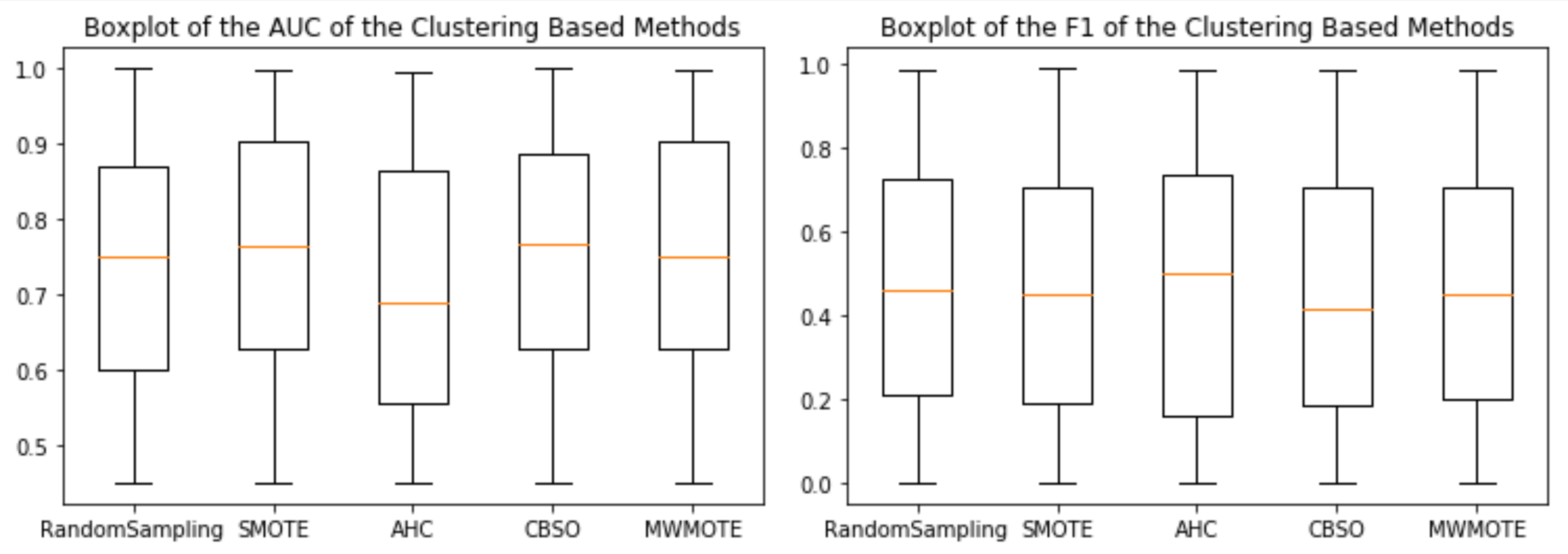
BOXPLOT OF RANGE RESTRICTED METHODS





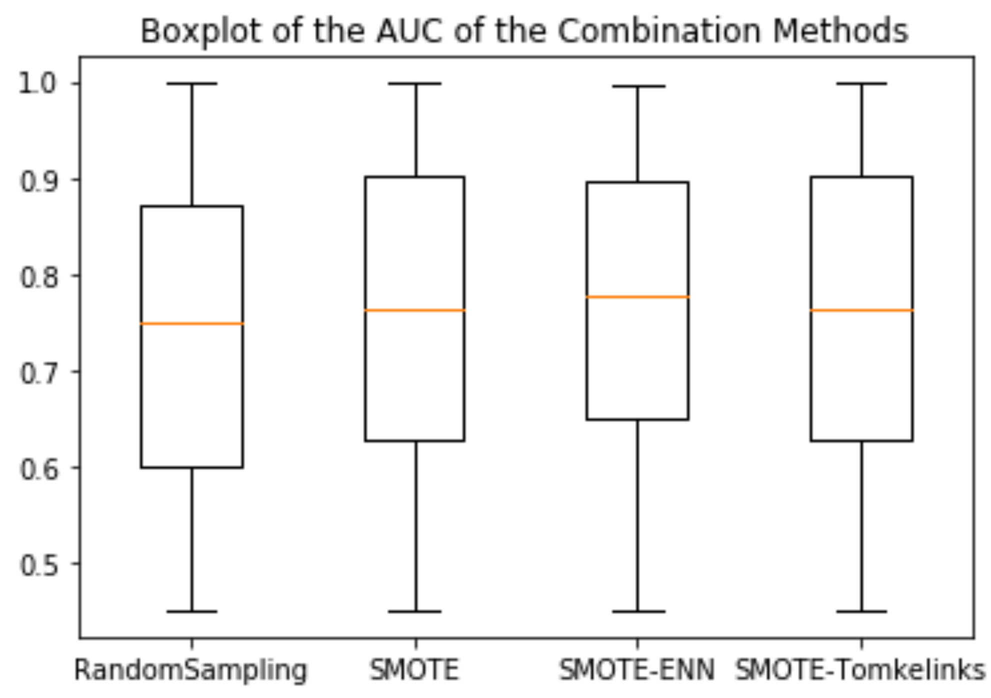
Appendix

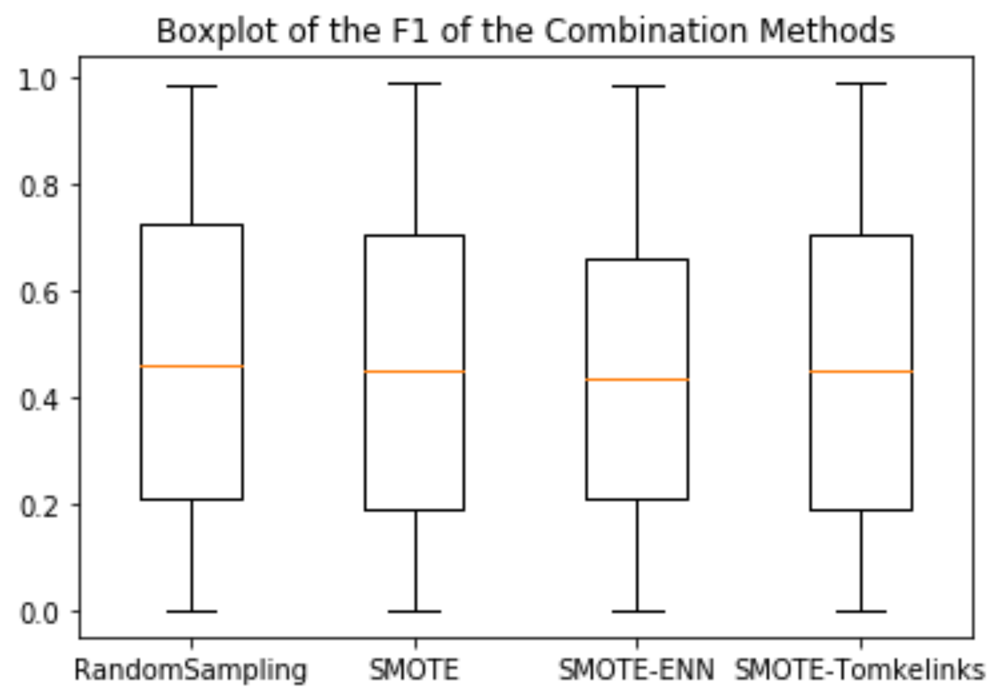
BOXPLOT OF CLUSTER BASED METHODS



Appendix

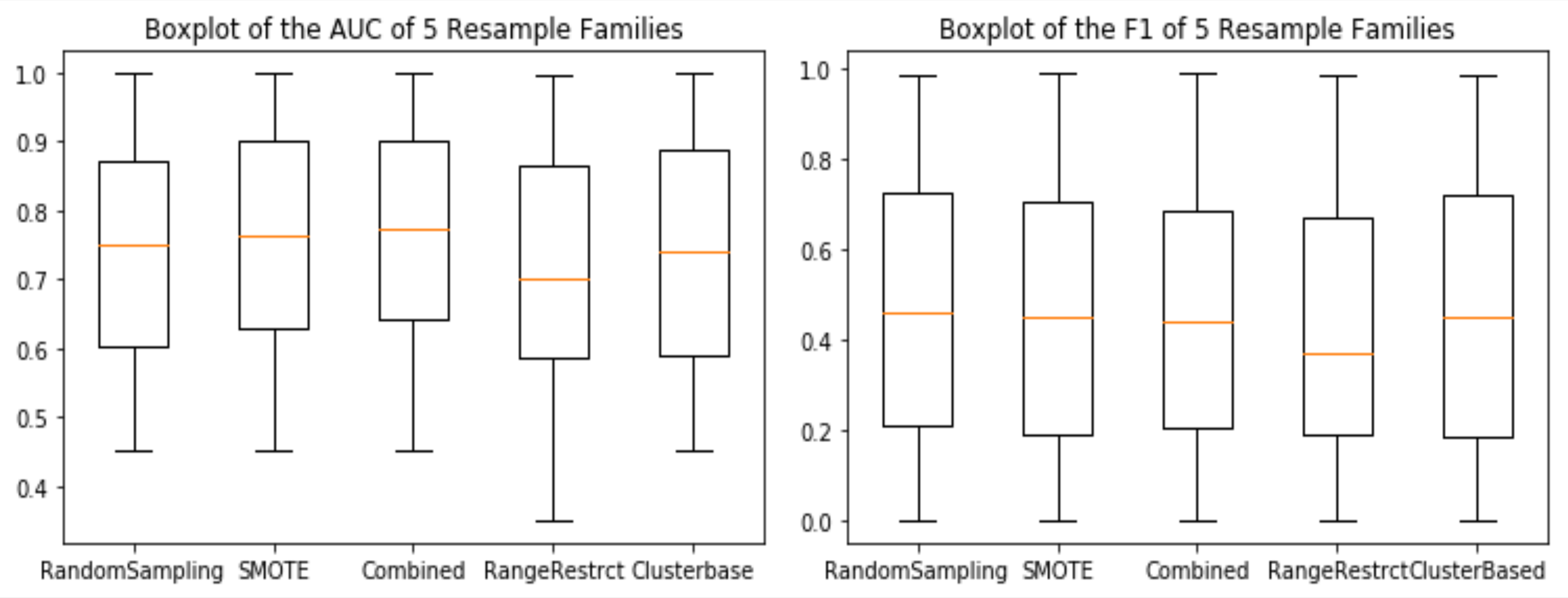
BOXPLOT OF COMBINING OVERSAMPLING AND UNDERSAMPLING METHODS





Appendix

BOXPLOT OF COMPARING THE 3 FAMILIES OF SMOTE EXTENSIONS



Appendix

SMOTE EXTENSIONS ON UNSTRUCTURED DATA