# Efficient Evidence Accumulation Clustering for large datasets

**Diogo Silva[1], Helena Aidos[2] and Ana Fred[2]**

[1]Portuguese Air Force Academy, Sintra Portugal

[2]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal

dasilva@academiafa.edu.pt, {haidos, afred}@lx.it.pt

## INTRODUCTION

- EAC is a robust ensemble method but its computational complexity restricts its use to small datasets.
- We propose an optimized implementation of the different EAC steps for faster execution and decreased memory usage.

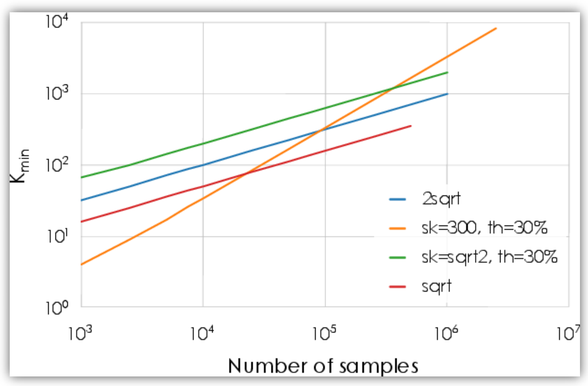
## VALIDATION AND SPEED-UP

- The clustering accuracy of the optimized version relative to the original on several small benchmark datasets is negligible.
- Speed-up over the original version on small datasets varied between 6 and 200 on the different EAC phases.

## RULES

Four rules for the minimum and maximum number of clusters of the ensemble were tested.

Evolution of $K_{min}$ with different rules




## PRODUCTION OF ENSEMBLE

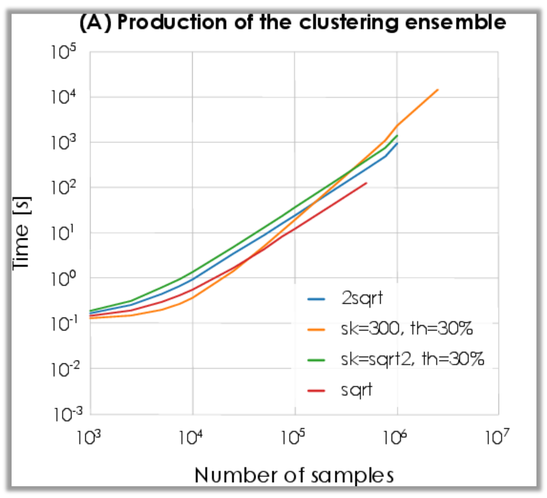
DATASET

**Challenge**
Fast generation of ensemble.

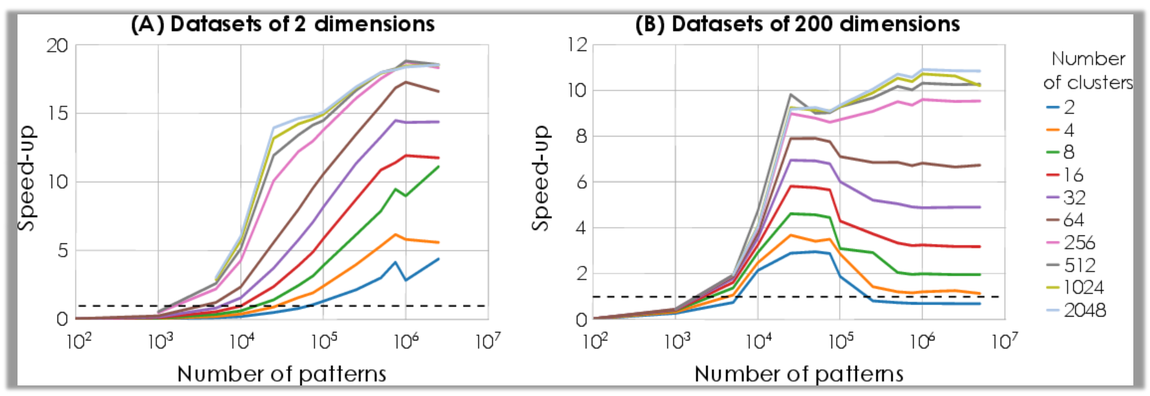
Partition *1*

Partition *P*

**Solution**
Parallel GPU K-Means




GPU K-Means speed-up

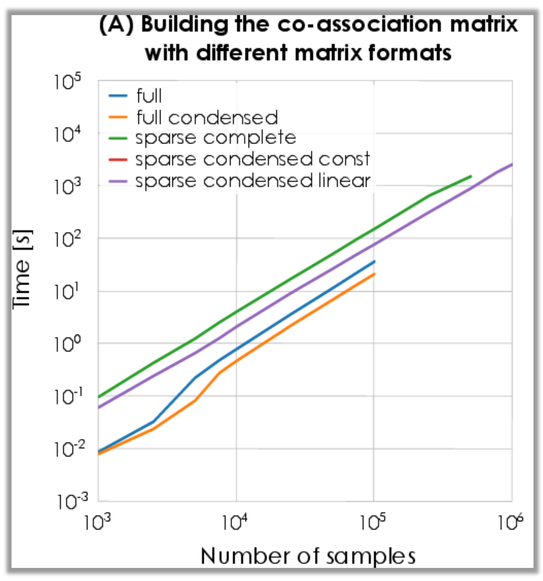


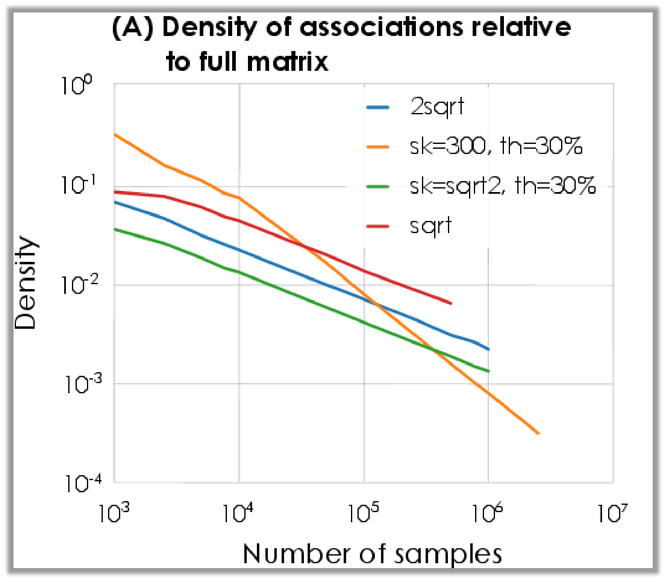

## COMBINATION OF PARTITIONS

**Challenge**
$O(n^2)$ space complexity

Co-association matrix

**Solution**
CSR sparse matrix with optimized building



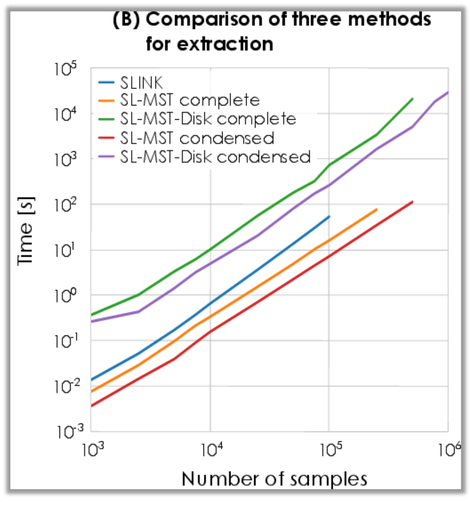




## RECOVERY OF FINAL PARTITION

**Challenge**
$O(n^2)$ space complexity

Single-Link (SL)

**Solution**
MST based SL
MST disk-based SL




## CONCLUSIONS

- EAC is now applicable to a wider spectrum of datasets.
- Speed-up from 6 to 200 compared to original implementation on the different phases.