Density based downsampling of cytometry data and clinical outcome prediction using clinical data

STU FIIT

Ing. Martin Nemček

doc. RNDr. Mária Lucká, PhD.

Problem

- Identification of cellular populations is one of the first and important steps in analysis of cytometry data which contain millions of events.
- To identify rare cell populations a density-based downsampling is performed.
- Using stochastic approaches to make algorithms usable on cytometry datasets in real-time render results irreproducible.
- Predicting clinical outcome using extracted features from cytometry data as well a clinical data enable discovering of novel relations.

Downsampling results

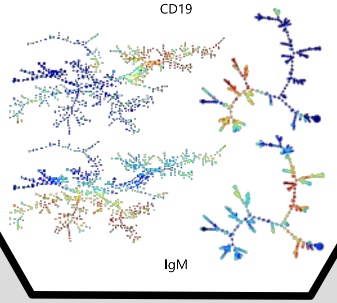
- The algorithm was compared with one of the most used software tools for cytometry data analysis – SPADE.
- Comparison was performed on multiple dataset of varying sizes for both tasks of density calculation and density based downsampling with and without the widely used arcsinh transformation.
- On each dataset the proposed algorithm achieved significant improvements in time complexity on both tasks.

Density calculation

- The parallel space partitioning algorithm minimizes computations needed to calculate a density of a point in a space.
- The space is partitioned based on ε-neighborhood of the points and efficiently represented by tree structure.
- A weighted density is calculated to address the drawbacks of density calculation while providing more precise results.

Biological results

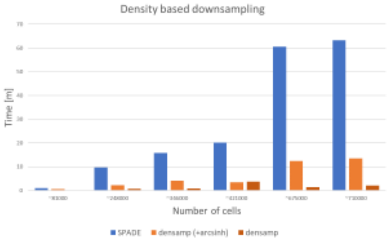
- The biological correctness of the results was approved by the domain experts.



Density based downsampling

- In each iteration of iterative density based downsampling a size of the ε-neighborhood is adjusted and the size of the dataset is reduced to sequentially reach the result.
- The iterative approach is based on two key observations.
- Density calculation of points in space partitioned by relatively small ε-neighborhood is very fast.
- Iterative reduction of dataset while proportionally increasing the size of the ε-neighborhood results in improved time complexity of entire process.

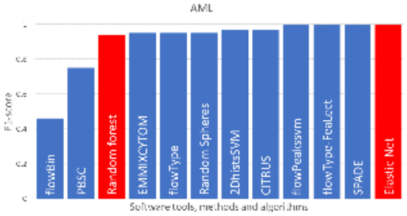
Comparison



Prediction

- Random forest and Elastic net models were build using extracted features from cytometry and clinical data.
- Patient's state and responses to specific treatments were predicted.
- The proposed method achieved the best results on AML dataset from FlowCAP-II competition.

Prediction results



Conclusion

- Novel approach to deterministic density based downsampling.
- Significant improvements in time complexity.
- Results approved by domain experts.
- Comparable one of the best prediction results.
- Cooperation with Slovak Academy of Sciences.