Levelwise Clustering under a Maximum SSE Constraint

Jeroen De Knijf Bart Goethals and Adriana Prado

ADReM
Computer Science and Mathematic Department
Antwerp University

Research Question

Question

Can we use techniques from frequent itemset mining to find an optimal clustering of points in Euclidean Space ?

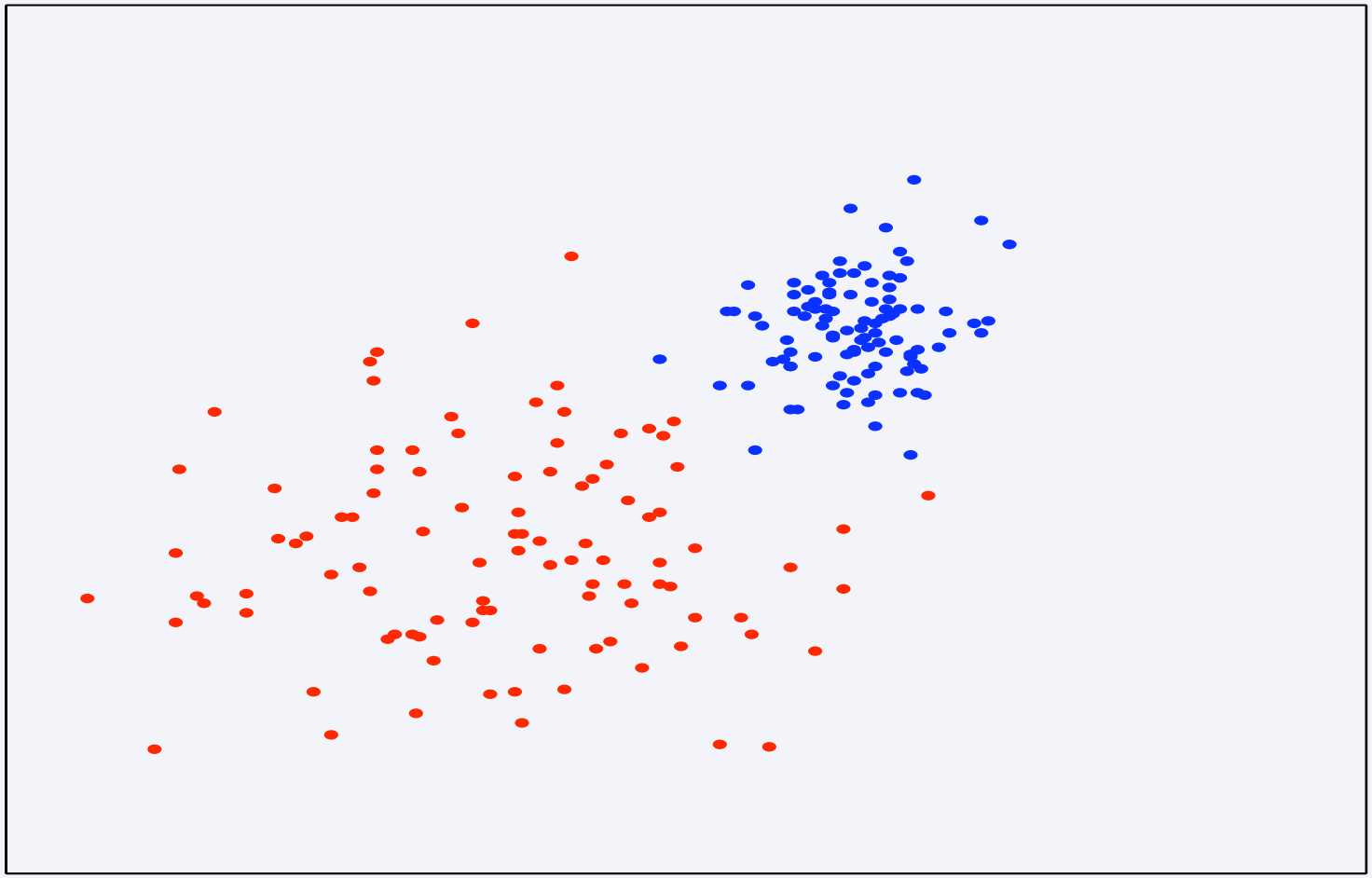
Introduction

- Set of points $P \subset \mathscr{R}^d$
- Sum of Squared Errors as optimization criterion.
- Sum of Squared Errors is monotone with respect to set inclusion.

Basic Algorithm

1. Derive all sets with an SSE value lower than MAXSSE (Apriori-style mining algorithm).
2. Use the derived sets to construct a partition of the data (greedy heuristic).
3. Create a new dataset by replacing the points in the clusters by its centroid. Continue with the first step.

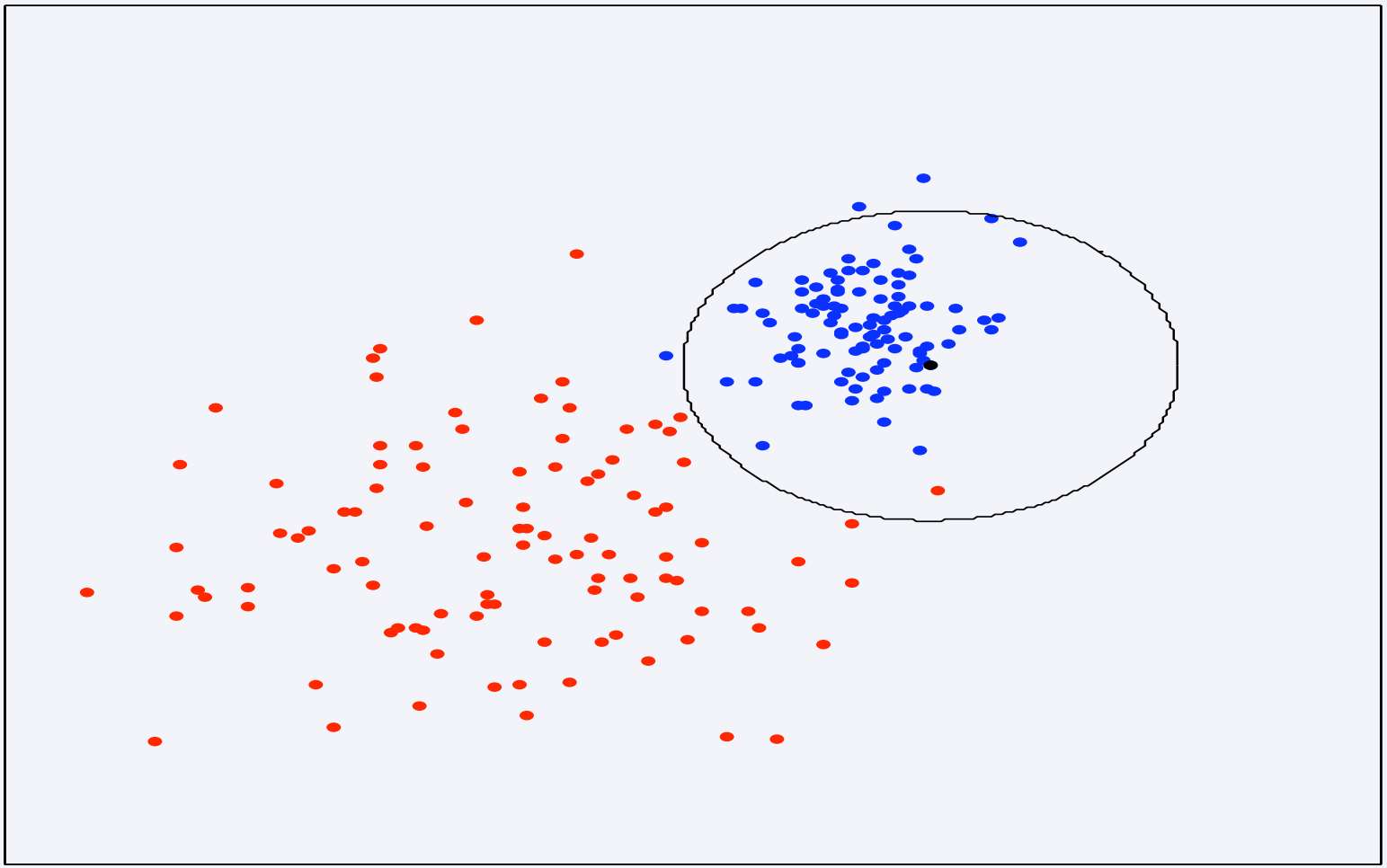
One max SSE value for the whole input space ?



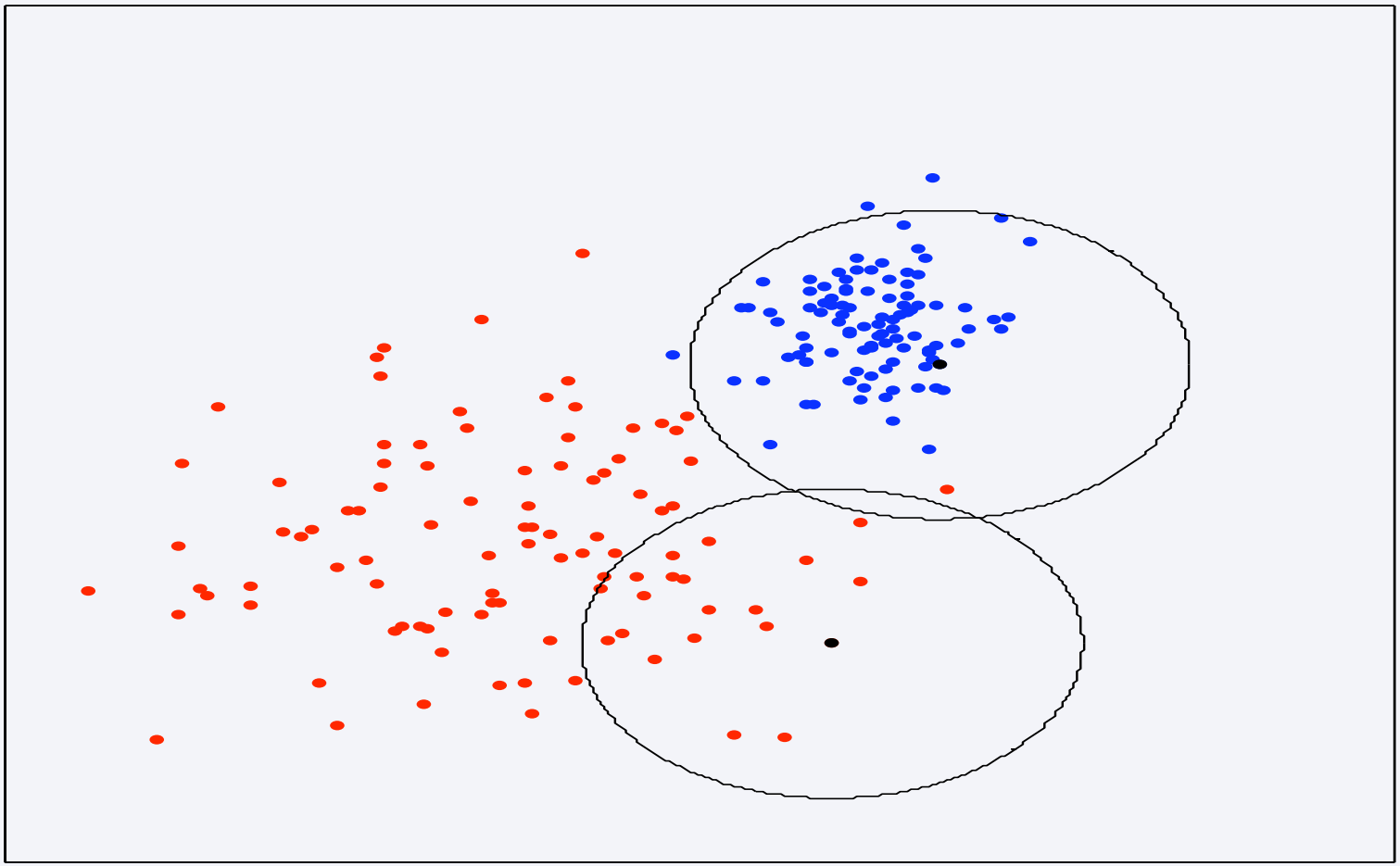
Mining Algorithm

- Perform $|P|$ mining algorithms, one for each point $p \in P$ and its region.
- The MAXSSE value for the mining algorithm of point p is depended on the density of the region of p.

The region of a point



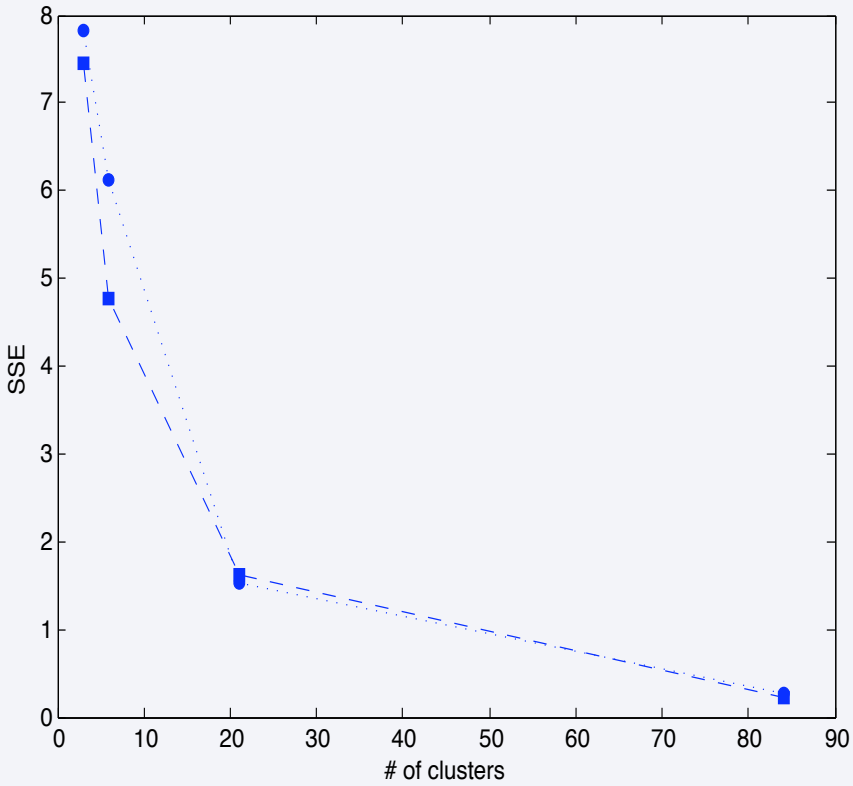
The region of a point

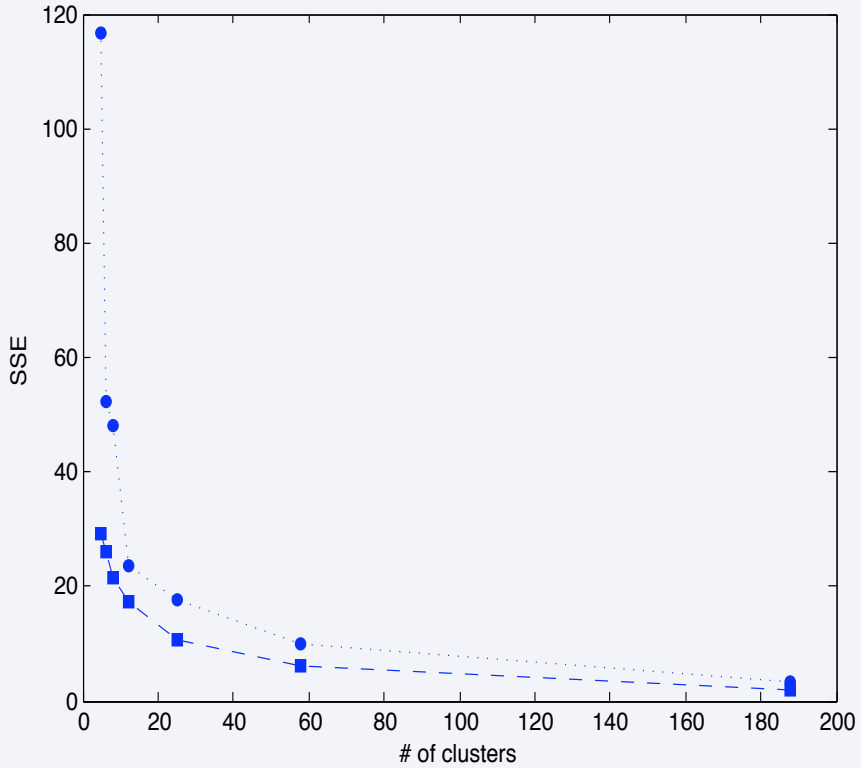


Dataset

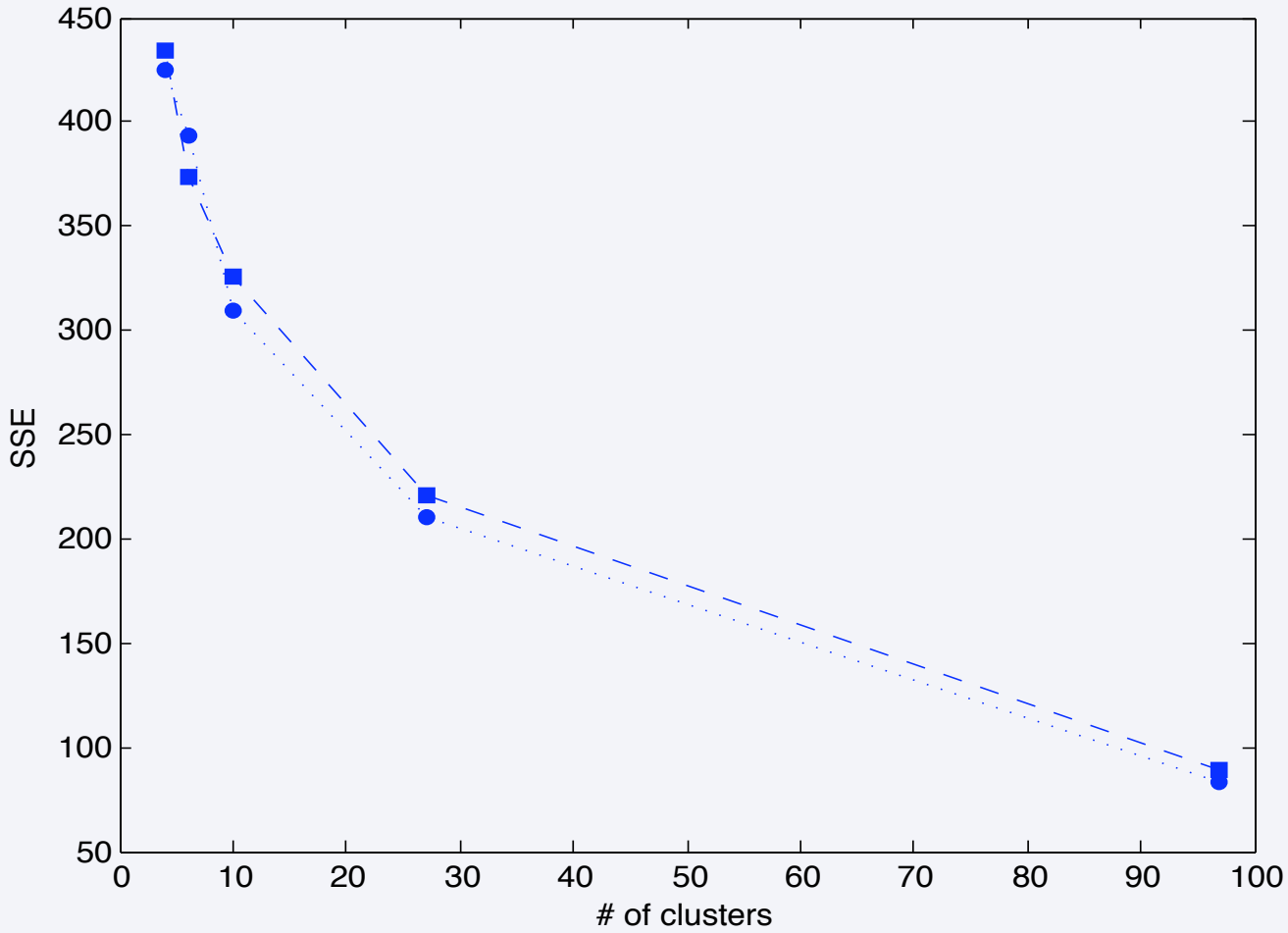
Dataset	#points	#dimensions	#classes
Iris	150	5	3
Ecoli	336	8	8
Sonar	208	61	2

Results on the Iris and Ecoli dataset.





Results on the Sonar dataset.



Conclusion

- Our approach is able to capture essential clusters in a dataset.
- The number of cluster obtained is unpredictable and rather sensitive to the input parameter.