Clustering Techniques: A comparison

Deepthi Hegde, Mathieu Gulliame-Bert, Kyle Miller, Artur Dubrawski

Carnegie Mellon University

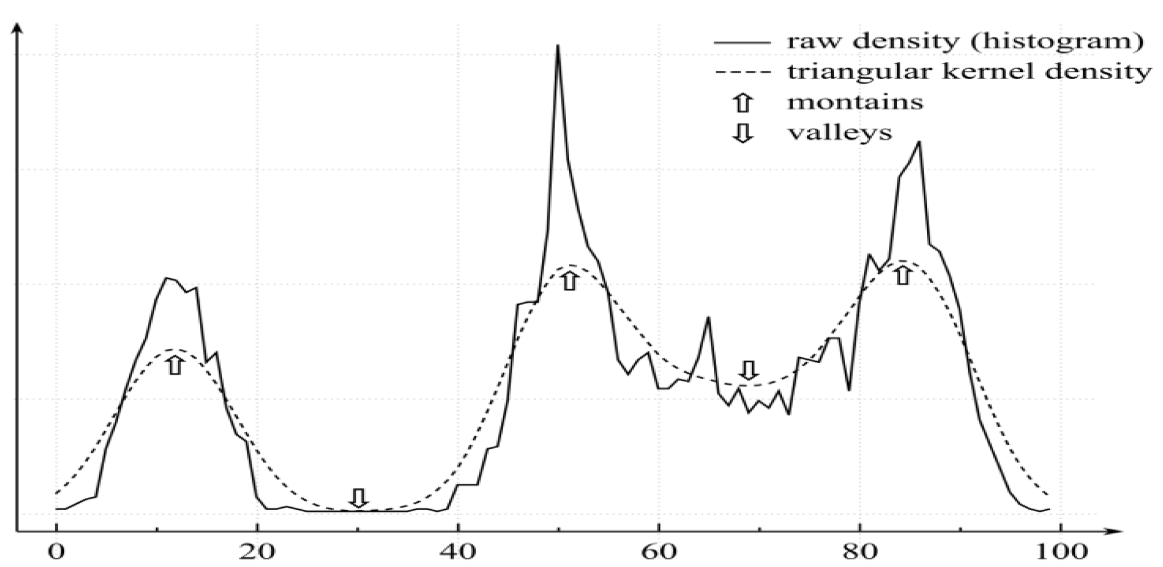
THE ROBOTICS INSTITUTE

Introduction

- Custering is the process of automatically finding groups of similar objects in data.
- Marginal clustering is clustering by detecting suitable planes that separate groups of similar data called clusters..

➢ We aim to compare and evaluate various marginal, distance based and density based clustering methods.

Illustration of peaks

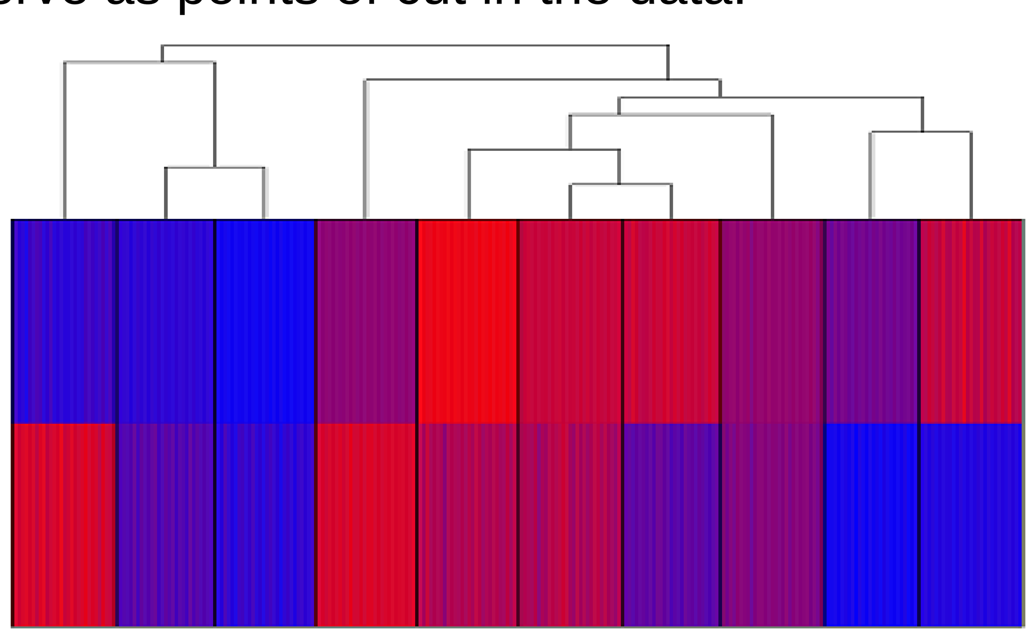


Peaks: Points of local maxima and minima
Mountain: Maximum peak
Valley: Minimum peak
Cut: a plane separating clusters

➢ A deep valley between 2 tall peaks indicates region of low density suitable for a cut.

Split Clustering

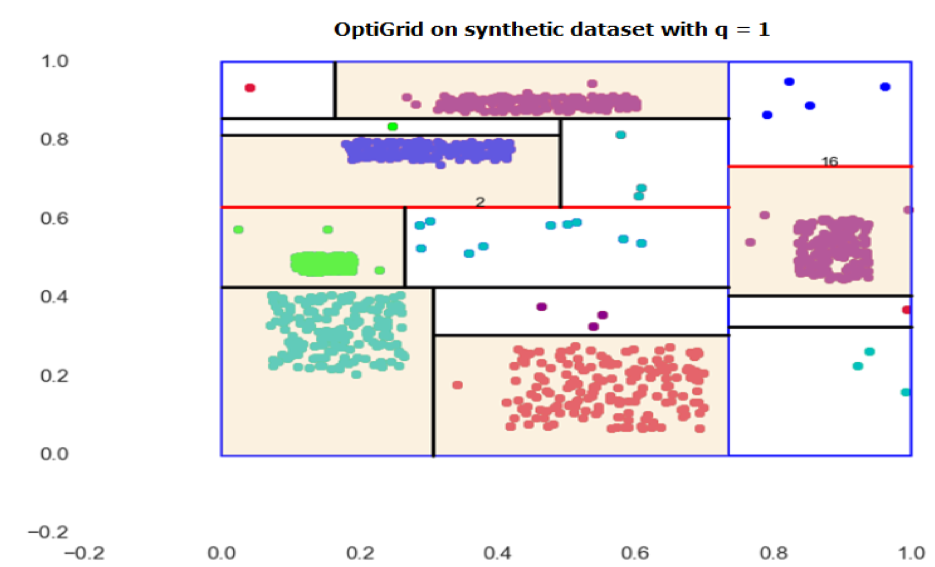
Split clustering aims at uncovering clusters in high dimensional data with structure in lower dimensions. Hyperplanes that separate the data into groups are found recursively. Valleys serve as points of cut in the data.



Dendrogram for split clustering

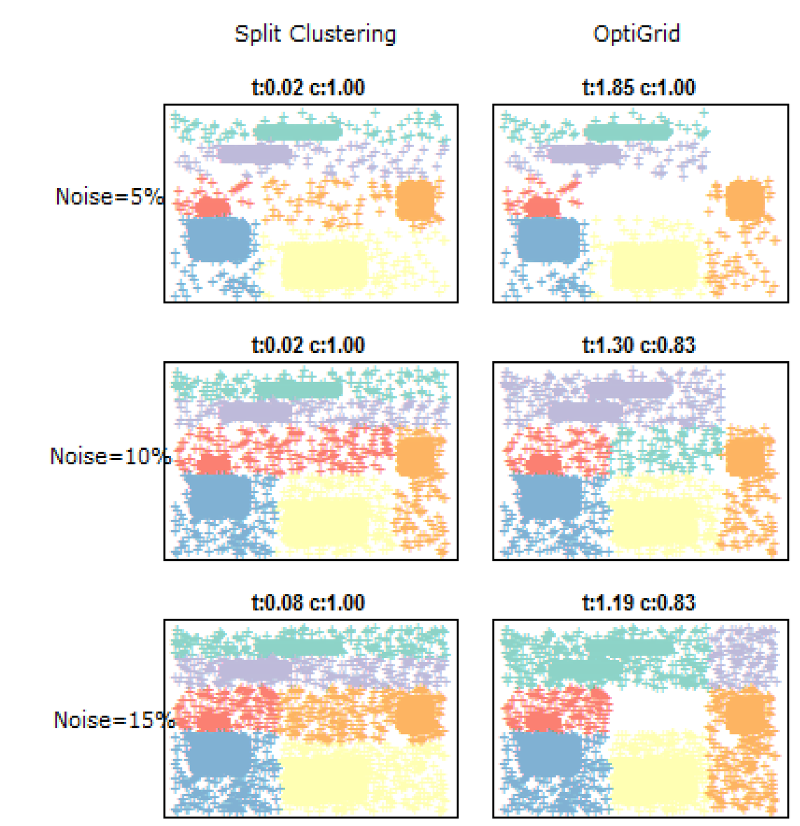
OptiGrid

OptiGrid is a grid partitioning clustering technique that recursively finds optimal cutting planes to find clusters effectively. The algorithm cuts the dataset into grids and recursively finds more cuts (if possible) within each of the dense grids. The cuts are identified based on the density curve of each dimension of the dataset independently. The top q cuts across all dimensions are chosen based on the density score.

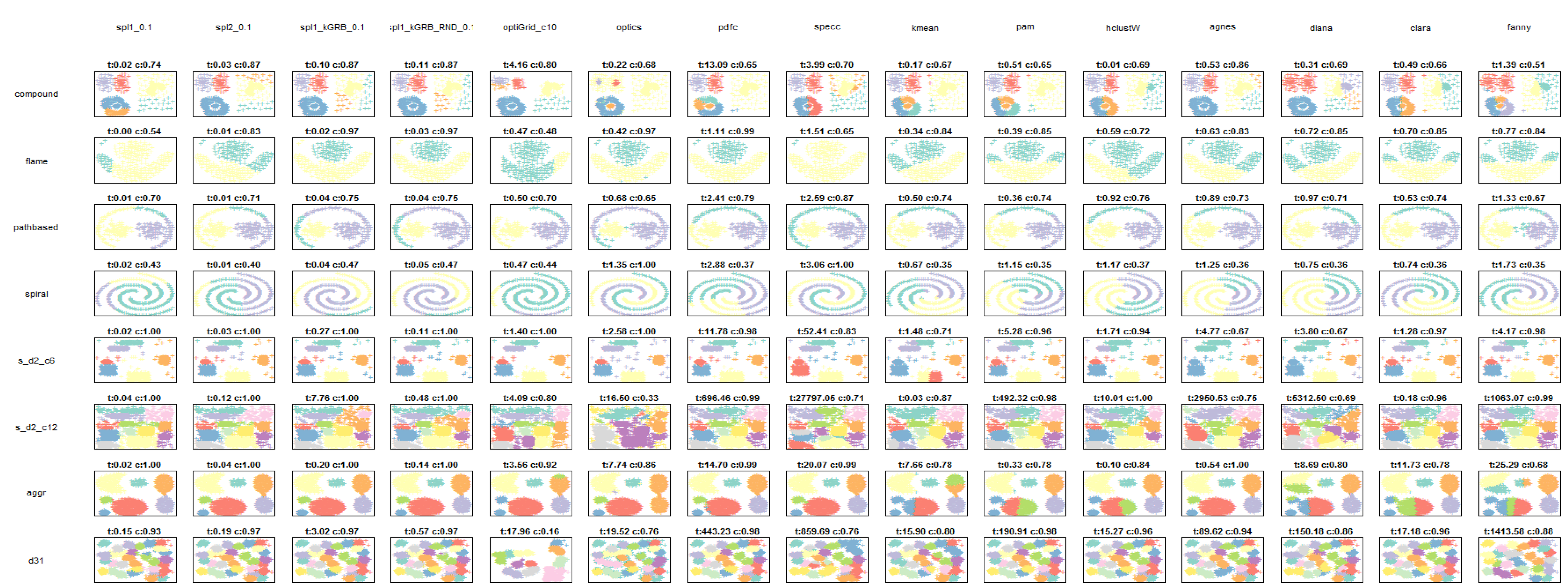


Split vs OptiGrid: Noise comparison

Split clustering is robust to noise. The structure of detected clusters is not hampered by the presence of noise points in the data.



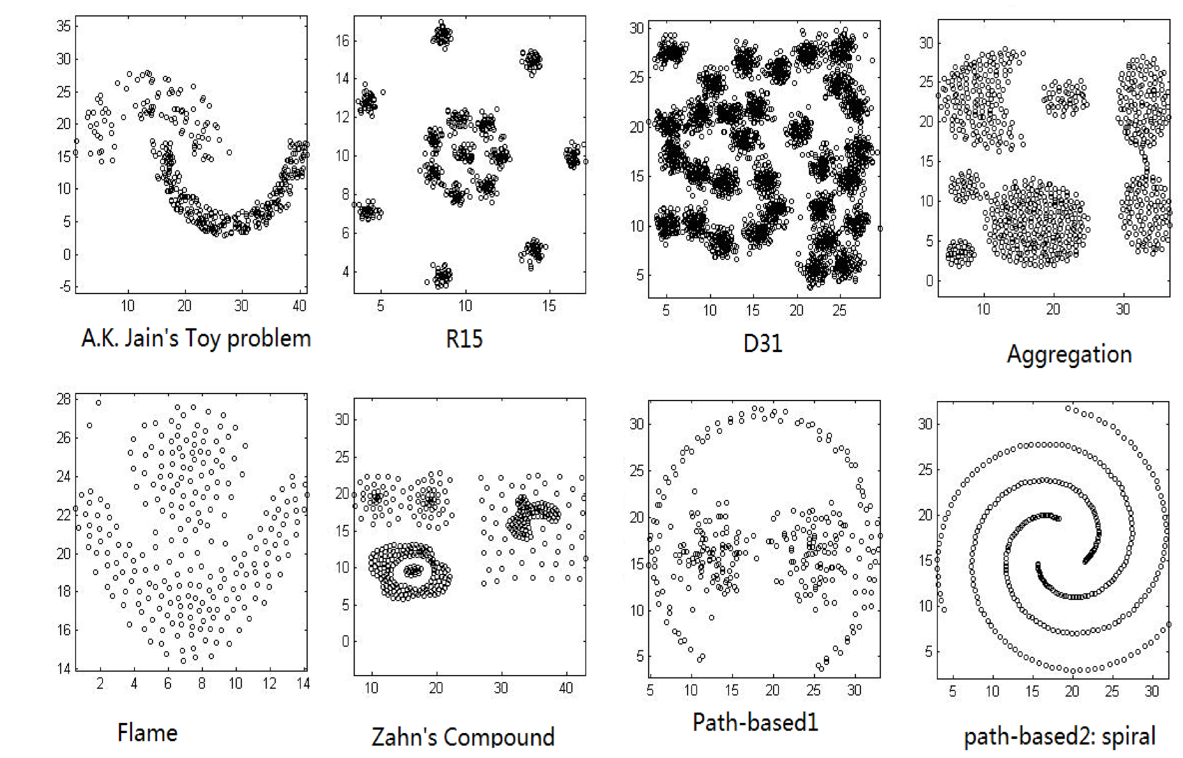
Split vs Others



Results

Algo\Data	Synthetic	Synthetic	Synthetic	Compound	Flame	Pathbased	Spiral	Aggregate	Iris	Wine	Breast	Average
split 1D	1	1	0.833333333	0.736842105	0.5375	0.696666667	0.426282	0.996193	0.96	0.606742	0.866432	0.7872719
split 2D	1	1	1	0.86716792	0.829167	0.706666667	0.403846	0.998731	0.96	0.44382	0.653779	0.8057434
split GRB	1	1	0.833333333	0.86716792	0.975	0.746666667	0.474359	0.996193	0.886667	0.539326	0.838313	0.8324568
Split GRB RND	1	1	1	0.86716792	0.975	0.746666667	0.474359	0.996193	0.826667	0.539326	0.838313	0.8421538
optiGrid	1	0.801920768	1	0.804511278	0.479167	0.696666667	0.442308	0.923858	0.846667	0.38764	0.637961	0.7291545
optics	1	0.333333333	1	0.681704261	0.975	0.65	1	0.859137	0.793333	0.578652	0.720562	0.7810656
pdfc	0.98293173	0.994297719	1	0.651629073	0.991667	0.79	0.365385	0.991117	0.893333	0.651685	0.727592	0.8217852
spectral	0.83333333	0.708583433	0.833333333	0.696741855	0.645833	0.87	1	0.993655	0.9	0.629213	0.862917	0.8157828
k-means	0.71084337	0.8737495	0.833333333	0.666666667	0.8375	0.743333333	0.346154	0.784264	0.893333	0.702247	0.85413	0.7495959
pam	0.96485944	0.982292917	1	0.646616541	0.85	0.74	0.349359	0.777919	0.893333	0.707865	0.86819	0.7982214
hclust	0.93674699	1	1	0.691729323	0.720833	0.76	0.371795	0.837563	0.893333	0.696629	0.778559	0.7897445
agnes	0.66666667	0.75	0.333333333	0.862155388	0.833333	0.73	0.358974	1	0.906667	0.61236	0.662566	0.7014596
diana	0.66666667	0.68597439	0.5	0.689223058	0.854167	0.713333333	0.358974	0.798223	0.88	0.52809	0.850615	0.6841152
clara	0.96787149	0.962685074	1	0.656641604	0.85	0.743333333	0.358974	0.784264	0.893333	0.707865	0.86819	0.799378
fanny	0.97991968	0.990896359	1	0.513784461	0.841667	0.666666667	0.352564	0.681472	0.913333	0.707865	0.887522	0.7759719
Clusters	6	12	6	6	2	3	3	7	3	3	2	NA
Rows	1046	10496	1046	399	240	300	312	788	150	178	569	NA
Dimension	2	2	4	2	2	2	2	2	4	13	30	NA

Versatility of Split Clustering



Split clustering can handle a wide range of datasets with different cluster shapes. While the 1D and 2D cuts can detect convex clusters, the kernel trick used in the algorithm works well with non convex clusters.

Conclusion

Split clustering algorithm is –

- Simple and easy to implement and interpret.
- Robust to noise
- Very low computation time
- Few parameters
- Capable of handling randomly shaped clusters

References

[1] *Alexander Hinneburg and Daniel A. Keim. Optimal gridclustering: Towards breaking the curse of dimensionality in high-dimensional clustering. pages 506–517. Morgan Kaufmann, 1999.*

Acknowledgements

Thanks to all the members of the Auton lab and RISS team for the great experience. Special thanks to Mrs. Sowmya Kamath and Prof. Ram Mohan Reddy for their support.