SUBCLU

Micah Nacht and Nick Spinale

DBSCAN (1996) - Background

Spatial database: represents objects in a geometric space

Spatial index: data structure for optimizing space-related queries

Want to cluster a spatial database supported by a spatial index using some notion of density

[demo]

DBSCAN (1996) - Definitions

- Core points
  - Points that have dense surroundings
- Density-reachability
  - All points in a dense region should be part of the same cluster
- Border points
  - density-reachable from core points, but not core points themselves
- Noise points
  - Points that are not in a cluster
- Cluster
  - All points density-reachable from a core point

Curse of Dimensionality

As dimensionality increases...

- Almost all pairs of points are equally far away from one another
- Almost any two vectors are almost orthogonal

Clusters and noise in even mildly irregular data become nearly impossible distinguish

SUBCLU (2004)

Use DBSCAN to find all clusters in each subspace

Monotonicity of density-connectivity [demo]

A-priori-like bottom-up greedy algorithm

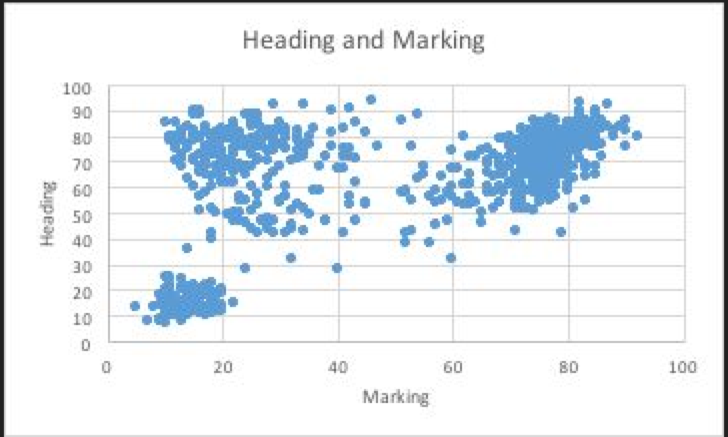
Find clusters hidden in subspaces [demo]

Clustering FIFA Players - Data

- Large dataset of player attributes from FIFA 17
  - Preferred positions
  - Physical characteristics
  - Soccer skills
- Took a subset
  - Random 10,000 players with single preferred positions
- Applicable for this algorithm
  - Each dimension has a meaning
  - 40 dimensions is not an unreasonable amount
  - Some interesting clusters may be hidden by noisy dimensions

Clustering FIFA Data - Results

- Found clusters that were highly correlated with position
- Results make sense
- Heading/Marking - GK in the bottom left
- Higher dimensional-subspaces identified ~90% CB cluster -
  - Heading, Strength, Standing Tackle, Skill Moves, Weight



Limitations

Data where different regions have different densities [demo]

Questions?

References

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Kailing, K., Kriegel, H. P., & Kröger, P. (2004, April). Density-connected subspace clustering for high-dimensional data. In Proceedings of the 2004 SIAM International Conference on Data Mining (pp. 246-256). Society for Industrial and Applied Mathematics.

Agarwal, Soumitra. "Complete FIFA 2017 Player Dataset (Global)." Kaggle.com. N.p., Mar. 2017. Web. 28 May 2017.