Paper Presentation:
Robust Distance-Based Clustering with Applications to Spatial Data Mining

by Vladimir Estivill-Castro and Michael Houle

to appear in 2000 in *Algorithmica*

Kevin Pulo

24 May 2000

Spatial Data Mining

- Large spatial datasets of points.
  - Too large to visualise.
  - Possibly higher dimensional.
- Want to find sets of similar points within the dataset.
- Exploratory analysis desires techniques which are automatic and fast.
  - Avoid human bias.
  - Fast response times aid the exploration.
- Most common application is 2 dimensional spatial data in Geographical Information Systems (GIS).

Clustering Methods

- Clustering is a key technique in Spatial Data Mining.
- Good quality clustering algorithms are usually expensive, ie. $O(n^2)$ or worse.
- k-MEANS is a simple clustering algorithm taking $O(n)$ time.
- k-MEDOIDS (*p-median*) is a more constrained version of k-MEANS, taking $\Omega(n^2)$ time.
- The paper presents a variant of a k-MEDOIDS heuristic, using the Delaunay triangulation to require only $O(n \log n)$ time.

Clustering

- k clusters, each represented by a *representative* point.

- The most appropriate number of clusters k may or may not be determined by the algorithm.

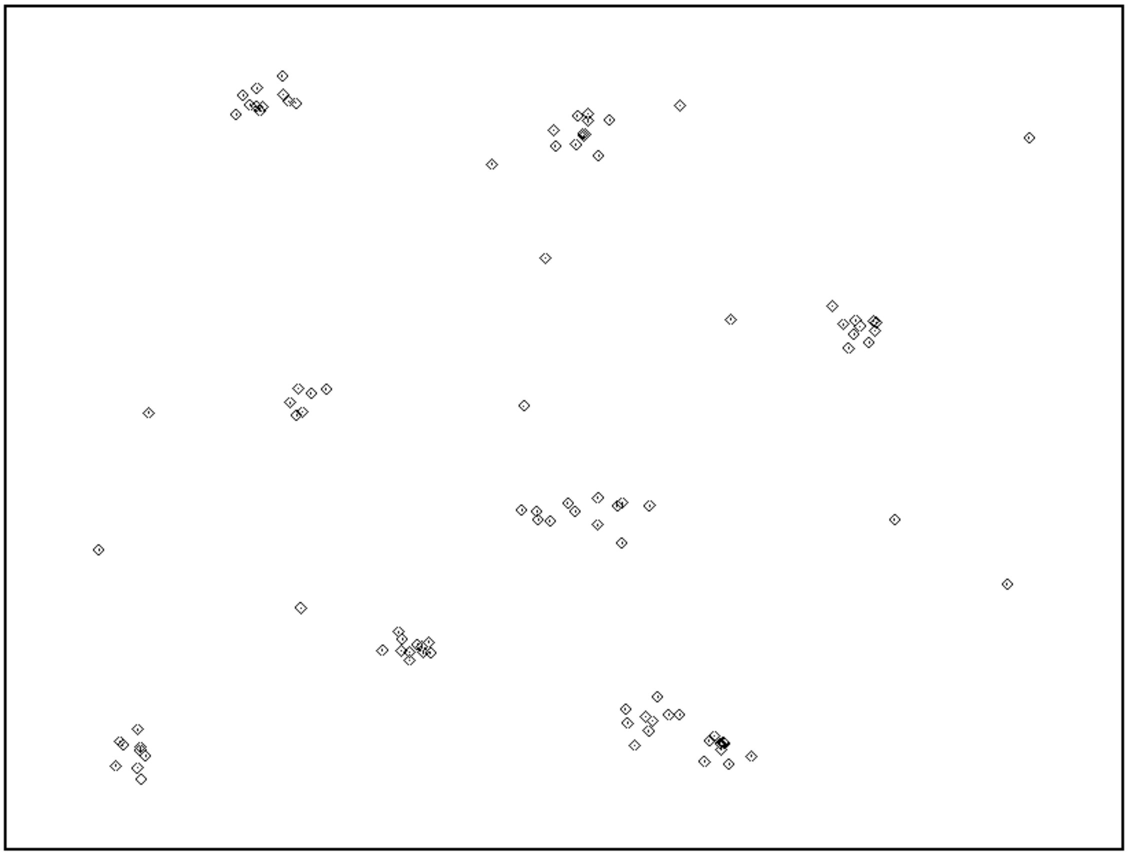
- Representative point may or may not be a data point of the dataset.

- Points need not belong to a cluster, eg. noise and outlier points.

- Clustering algorithms give approximate solutions to the problem of minimising the sum of the distances of each data point from their nearest representative point.

$$\text{minimise } M(C) = \sum_{i=0}^{n-1} d(s_i, \text{rep}[s_i, C])$$

Sample Dataset



Sample points lying in the 2 dimensional x-y plane.

k-Means

- Take an initial (perhaps random) clustering.
- For each cluster, find the new representative by computing the average of the cluster points.
- Using these new representatives, find a new clustering (ie. assign points to their nearest, possibly different, representative).
- Repeat these two steps until no change occurs.
- This is "non-combinatorial" reclassification, the "combinatorial" reclassification variant recomputes the clustering each time a new representative is computed.

k-Means: Advantages and Disadvantages

✔ Fast — $O(n)$ linear time.

✔ Simple.

✘ Requires k to be specified.

✘ Representative points may have no valid interpretation.

✘ Sensitive to noise, outliers and initial random clustering.

✘ Does not classify outliers.

✘ Statistically biased, often converges to a poor local optimum.

✘ Uses the square of the Euclidean metric.

k-Medoids

Similar to k-MEANS, but has the additional restriction that the representative points must be chosen from the set of data points.

The TB heuristic, by Teitz and Bart, 1968, is the best known benchmark:

- Maintains a selection of k representative points from the n data points.
- Considers the data points in a fixed circular ordering $(s_0, s_1, s_2, \ldots, s_{n-1})$.
- If s_i is a representative, it is skipped, otherwise, it considers interchanging s_i with s_j, for all possible representatives s_j.
- If any of these yield an improvement, the interchange is performed, and the search continues from s_{i+1}.
- The search stops with a local optimum when a full cycle of the data points gives no improvement.

The number of points considered for interchange per iteration is typically constant. However, all n points must be examined for termination.

The circular ordering of points balances the need to explore various interchanges against the greed of improving the solution quickly.

k-Medoids: Advantages and Disadvantages

✔ Representative points are taken from the dataset points.

✔ Uses the actual Euclidean metric.

✔ Higher quality local optima found, even with noise and outliers.

✘ Slow — $\Omega(n^2)$ superquadratic time.

✘ Requires k to be specified.

✘ Does not classify outliers.

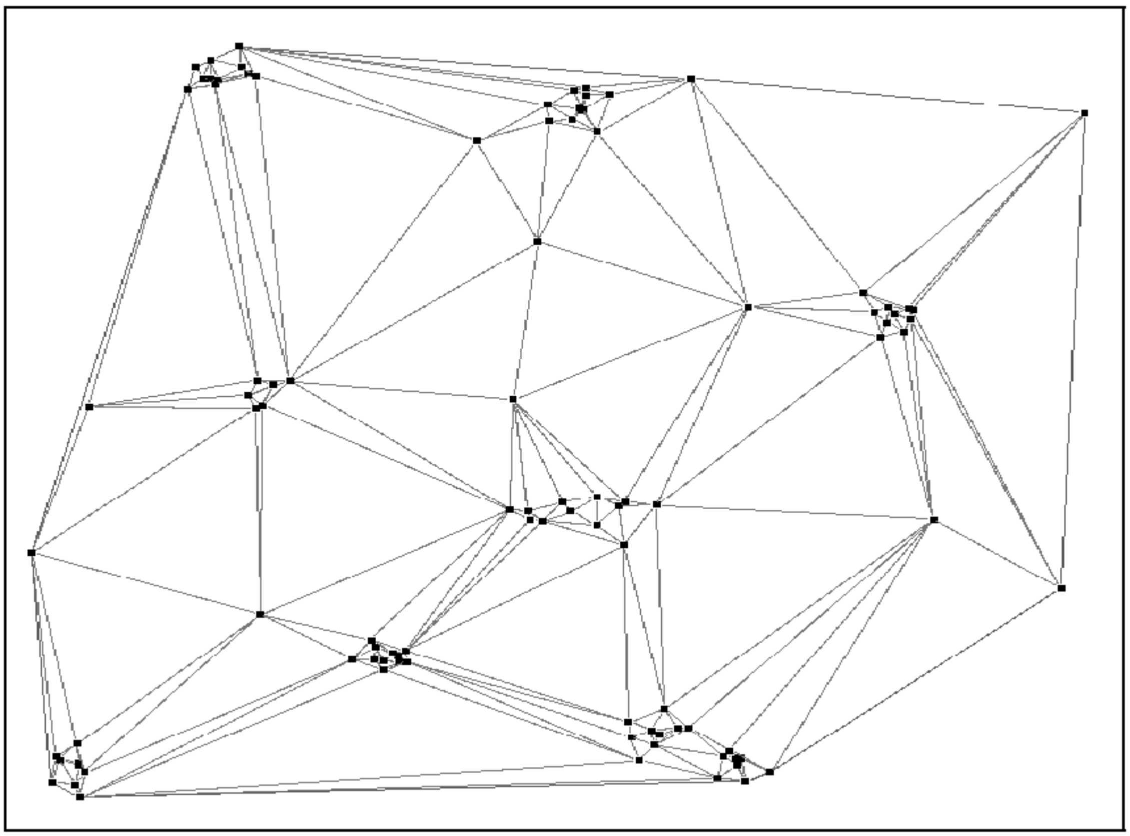
✘ Exact solution is NP-hard, ie. superpolynomial time complexity.

Delaunay Triangulation

Delaunay triangulations succinctly encapsulate the proximity information of a set of points.

- Points a and b are joined by the Delaunay edge ab iff they are "nearest neighbours".
- Delaunay edge $ab \Leftrightarrow \exists$ some circle through a and b containing no other points.
- For n data points, at most $3n - 6$ edges.
- Average number of neighbours of a point is less than 6.
- Minimum angle of all triangles is the maximum possible.
- Robust computation in $O(n \log n)$ time.
- u nearest neighbours of a point found in $O(u \log u)$ time.

Sample Dataset:
Delaunay Triangulation

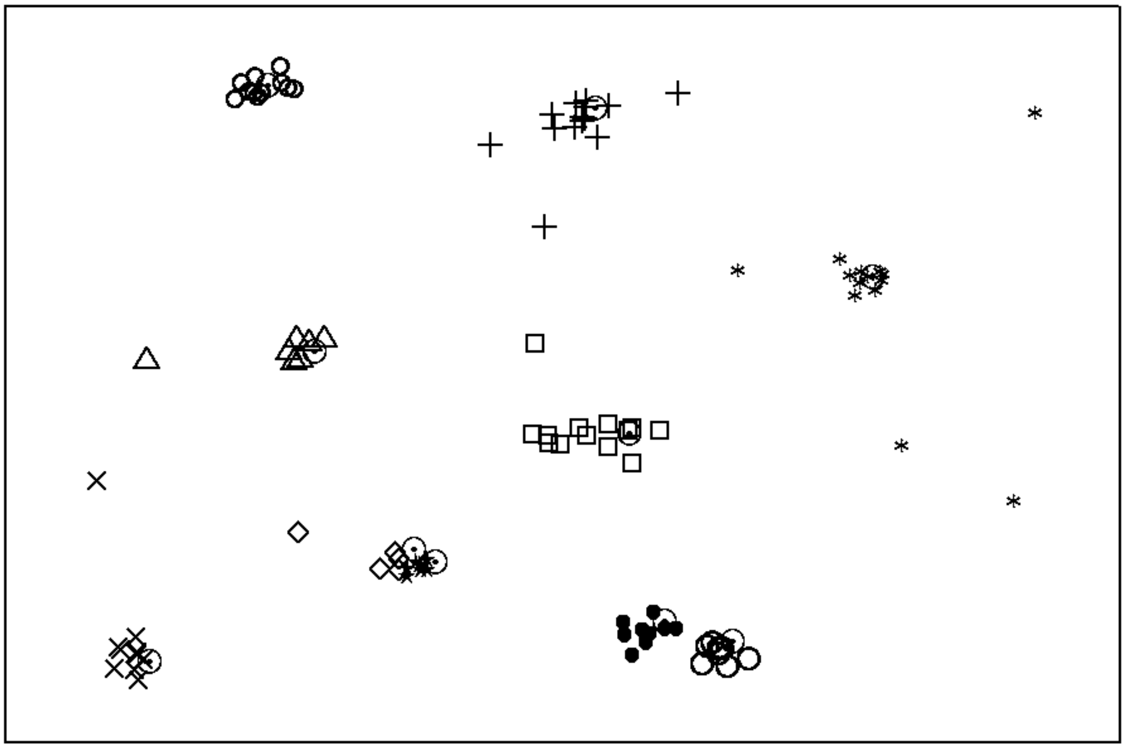


Sample points lying in the 2 dimensional x-y plane and their Delaunay triangulation.

k-Medoids Delaunay Variant

- Evaluating $M(C)$ exactly requires $O(n)$ time per interchange possibility.
- Greatest contribution to $M(C)$ is made by outliers (and initially there are more outlying representatives).
- Increase performance not by limiting the points examined by TB, but by approximating $M(C)$ with $M'(C)$.
- $M'(C)$ considers only the u nearest neighbours, precomputed for all points in $O(un \log n)$ time using the Delaunay triangulation.
- Evaluating $M'(C)$ takes only $O(uk)$ time per interchange possibility.
- If u is chosen to be $\Theta(\frac{\log n}{\log \log n})$, then the overall time bound simplifies to $O(n \log n)$. The choice of u determines a speed-quality tradeoff.

Sample Dataset: Clustering Results



The clustering of sample points produced by the algorithm.

k-Medoids Delaunay Variant: Advantages and Disadvantages

✔ Fast — $O(n \log n)$ subquadratic time.

✔ Representative points are taken from the dataset points.

✔ Uses the actual Euclidean metric.

✔ Discovers k automatically.

✘ Complicated.

✘ Does not classify outliers.

✘ Sensitive to initial representatives.

Finding an Initial Clustering

The following method is used to find the initial clustering in $O(n \log n)$ time.

- Put each data point into its own set.
- Sort the edges of the Delaunay triangulation by increasing length.
- For each edge, merge the sets connected by its endpoints.
- Stop when k sets of size at least ν (say, 2 or 3) remain.
- These sets form the initial clustering, representatives are chosen arbitrarily from them.

This technique maximises the minimum distance between initial clusters.

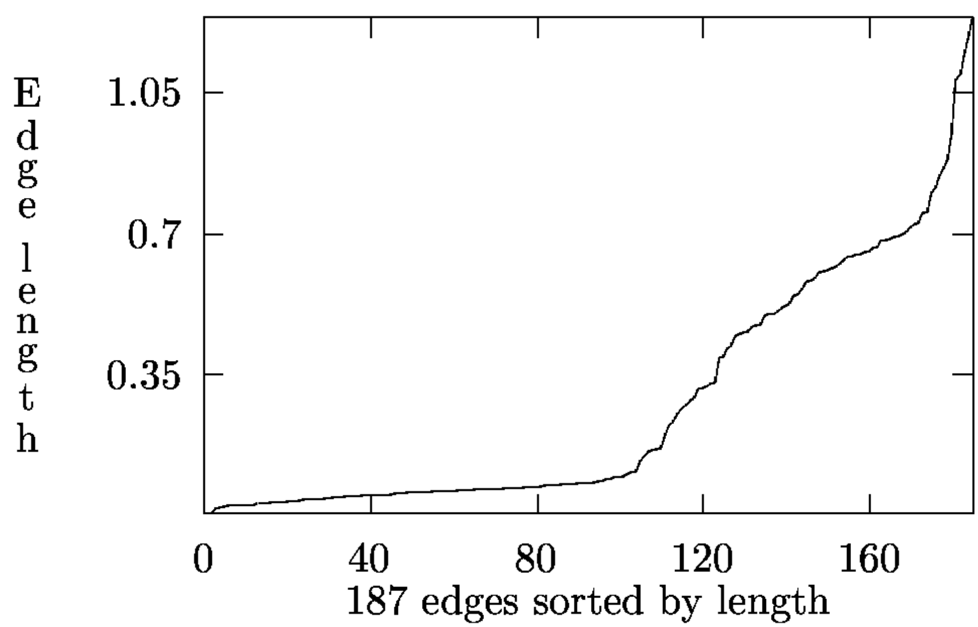
Finding k Automatically

This method can be adapted to find a suitable value of k for the dataset in $O(n \log n)$ time.

Consider the profile of sorted edge lengths.

Merge sets as before until just before the inflexion point, these are all the short edges which are "obviously" within clusters.

The number of sets of size at least ν is then taken as k, and the initial representatives chosen from these ν sets.



Comparison with Other Subquadratic Algorithms

- DBSCAN (1996) — $\Theta(n \log n)$ expected

  ✔ Finds clusters of arbitrary shape, not just convex.
  ✔ Good quality clusterings produced.

  ✘ Requires user assistance in removing outliers.

- STING (1997) — $O(n)$ construction, $\Theta(n)$ queries

  ✔ After initial processing, clustering time is sublinear.
  ✔ Structure similar to multidimensional database, thus suited to SQL-type queries.

  ✘ Grid size is hard to determine.
  ✘ Assumes mixture model of data, ie. assumes existance of probability distribution parameters.
  ✘ Scales poorly with the number of dimensions.
  ✘ Finds clusters which are orthogonal polygons.
  ✘ Gives poor, non-robust clustering results.

Experiments

- Clustered dataset generator program
  - Additive noise (random points through the dataset).
  - Multiplicative noise (random perturbation of points).
  - Guarantees two close clusters, relative to point spacings within clusters.

- Compared k-Means (random and Delaunay-based initialisations), original TB (k-Medoids), modified TB (both given k and not given k).
  - k-Means only competitive when no noise.
  - Modified TB competitive with original TB when no multiplicative noise.
  - With multiplicative noise, modified TB worse than original TB, but still better than the Delaunay-based initialisation of k-Means.
  - Modified TB, when not given k, performs at least as well as the Delaunay-based initialisation of k-Means.

Blank Slide