Theta A

Pragat Wagle

October 2021

1 Theta A

Clustering problems have been tackled using dimensionality reduction, density estimation, probabilistic methods, spectral methods, and distance based methods. Distance based are the most widely used methods but depend on the the fact that certain information is provided. Theta A uses the distance based approach by using a distance threshold with out any prior experience. The focus of ThetaA is to predict the number of clusters based on the the distance threshold rather than a fixed limit, with its focus the scientific domain where we have prior knowledge of the clusters dimensions we intend to find eg. cells and atoms of microorganisms.

2 Theory

2.1 Definitions

θ sparse consist of datasets that have a higher cluster distance than θ where θ dense have a distance less that θ

2.2 Theorem 3.1

Give clustering with clusters C_1, C_2, C_3, C_k, if no pairs exists in Ci to Cj, where i $\neq$ j have a distance less that the threshold θ, then algorithm 1 will identify the clusters with threshold θ at exactly N steps and order will not effect the final outcome. This will circumvent the issue of data ordering as in existing threshold based clustering clusters are branched in the sequence they appear which can issues. So this algorithm aims to improve accuracy while managing the issue with ordering.

Proof of Theorem
Distance are computed inside and between clusters. The distances inside a cluster ¡ θ while between ¿ θ. This is is so that no clusters are created between clusters given there are no pairs of samples xi, xj from the Ci to Cj that have a distance < 0 where C is a cluster. There exists no point from ci to cj where the point is $< \theta$, meaning the distance between any point sample with each respective cluster is $> \theta$.
Lemma
If the algorithm is repeated with random uniform samplings and the labels do not change it is sparse. The distance is greater than the threshold which means chance of label changing would be lower.
Proof of Lemma
By changing order the chance of choosing points between clusters with a distance ¡ θ increases and if that does not happen it is sparse. If the distance between clusters get smaller than this presents a bigger challenge as the chance

of have two points, both in their respective cluster, with a distance , θ increase thus algorithm 2 is presented which add to itself, an inverse proportionality of the number of orderings to the distance between the clusters. In cases like this reorder is important. Algorithm 1 and 2 solve linearly separable clustering problems while Algorithm 3 solves non-linear clustering problems.

3 Algorithms

3.1 ThetA Sparse Grouping (TSG)

Lower time complexity than Kmean++ Lloyds. Low space complexity as it just stores centroids. Input is the data set X with a size of N and the threshold θ. Time complexity is dependent on the size of X. Worst is that all points belong to their own clusters creating all singleton clusters, which is a time complexity of of $O((N^2 + N - 2)/2)$ and best being O(N) when there is a single cluster. Samples are inserted into new clusters if they are far from existing clusters otherwise to the closest cluster available that meets sample ¡ θ for that cluster.

3.2 ThetA Dense Grouping

Lower time complexity than Kmean++ Lloyds but slower when iteration term is significantly higher than Lloy's, with a complexity of $O(I(N^2 + N - 2)/2)$. Space Complexity is low as it just to store centroids. TDG is run with number of iterations which shuffles the data with each iteration and then runs TSD on all collected centroids and then reassigns all points to the final centroids. Due to the shuffling the assumption is that centroids will be sparse as the number of centroids is less than the number of samples. A dense setting would imply that in a certain shuffle a data point cluster A was assigned to cluster B. With each the difficulty decreases as centers begin to reside towards the center of the cluster. This is done through calculations using true positives and false positives. It can be see that difficulty decrease by as long as m $< \frac{4}{7}$, which shows the cases where the center is within threshold θ. n

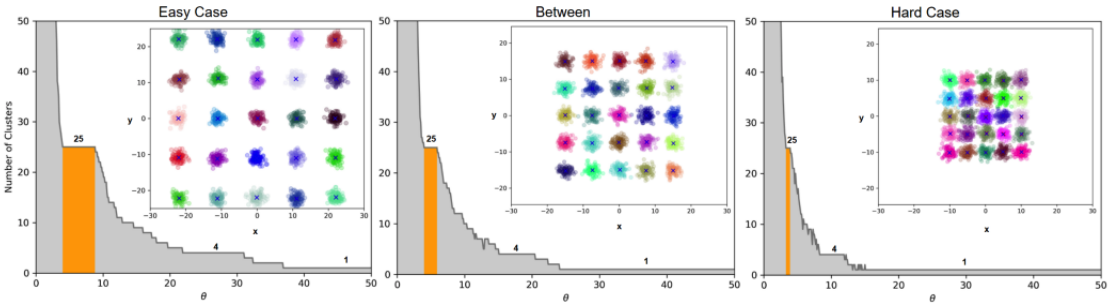
3.3 ThetA Nonlinear Chaining (TNC)

Algorithm 3 is an application example of how we can use TDG for nonlinear clustering problems. The idea is that first we start with TDG to produce small clusters and then we chain clusters together that are close to each other.

4 Cases

4.1 Summary Of Cases

Cases consist of the Easy and Hard case. The easy experimentation was done with 100 clusters with each centroid wide enough to be considered θ sparse, with a euclidean distance of 10.
The hard case was run with 100 clusters and a euclidean of 5 , with 50 shuffles and a θ of 3.6. This leads to clusters that are closer each other which can create problems associated with a dense clustering mentioned above.



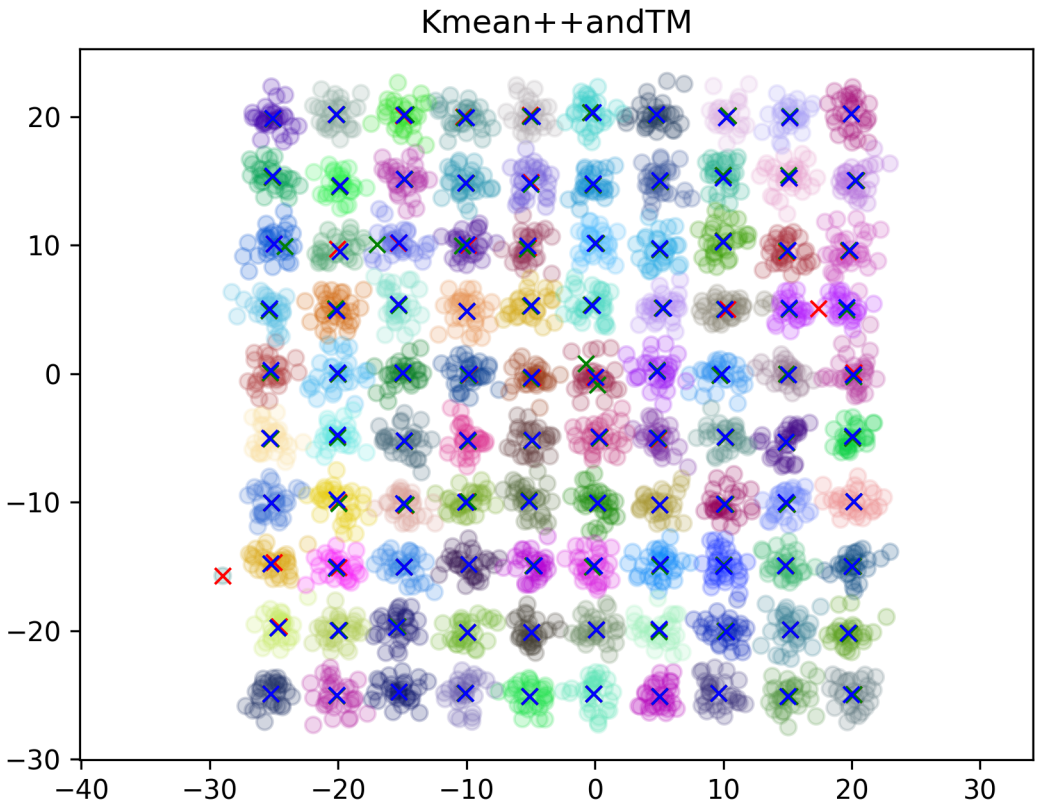
This figure was pulled from the article ThetA - fast and robust clustering via a distance parameter

5 Experiments

5.1 Summary Of Experiments

Through the various experiments we can see that ThetA outperformed. With higher dimensions convergence to a local optima rather than global optima can be seen in various existing methods but ThetA outperformed while other fluctuated or decreased. With sparser, less dense data optimal θ decrease. In the case of a extreme dense problem variations in threshold θ were added while lowering them and cluster sizes studied. We would expect with larger threshold for there to be a less clusters.

This is a test run of both Kmeans++ and ThetA overlapped and it can be seen that ThetA where overlap can be visualized is more accurate. Red represents ThetA, green represents Kmeans++ and blue the initial known centroids.


Kmean++andTM
20
10
0
−10
−20
−30
−40
−30
−20
−10
0
10
20
30