Approximation in Large Summation Problems using Sampling

Md Enayat Ullah, Indian Institute of Technology Kanpur

Mentor - Dr. Geoffrey Fox, Indiana University

Introduction

Bio-sequencing techniques have led to collection of genomic data, which usually is high dimensional and calls for the application of data mining algorithms. DACIDR[YR] is an application that generates robust clustering and visualization results on millions of sequences.It employs Multidimensional Scaling (MDS) to reduce the dimension of original data and pairwise clustering to classify the data. MDS is an umbrella term for the set of statistics techniques used for dimensionality reduction. The object is to construct such a mapping so as to reduce the dimensions to target dimension space while preserving the correlations in Euclidean distance in both the spaces. It is a non-linear optimization problem, which is solved iteratively by Deterministic Annealing, an EM algorithm which finds the global optima of an optimization process by adding a computational temperature to the target object function.

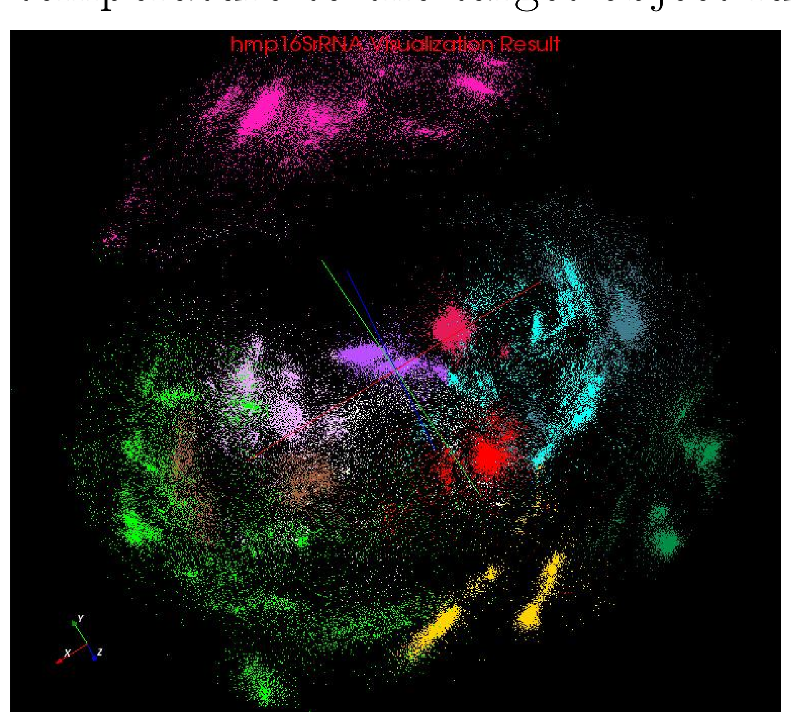

Figure 1: Clusters Visualized
Source: DACIDR[YR]

The object function in DA is essentially an n-body simulation [GM00], which takes a time of $O(n^2)$ to compute. Various algorithms have been studied in the past to improve the time complexity from $O(n^2)$ to $O(nlogn)$ or $O(n)$. Treecode methods like Barnes-Hut Simulation[BH86] reduces the complexity to $O(nlogn)$. Fast-multipole methods[GR87] take advantage of the fact that the multipole-expanded forces from distant particles are similar for particles close to each other and reduces the complexity to $O(n)$. ASKIT paper[WBMB] incorporates sampling ideas along with the existing treecode methods to improve the efficiency of a kernel summation problem.

Methodology

The summation which is to be approximated expands into $\binom{n}{2}$ terms and for every point in the 3 dimensional space, we require to compute the interaction with every other point. It is conjectured that only a fraction of terms in the summation can approximate to the exact summation, and so taking a sample from the not-so-important terms and weighting to generalize over all such points can theoretically decrease the computational time. The proposed method to group all the data point into near and far constructs a topological embedding of near and far regions which would represent a hollow sphere(the mapped dimensionality is 3), the thickness of the spherical surface(shell) is representative of the region, and all the points lying on the thick shell are to be approximated. The thickness varies the amount of approximation we intend to make.

Formulae for Stress

$$\sigma(X_T) = \sum_{i<j<=n} w_{ij}(d_{ij}(X_T) - \delta_{ij})^2$$

- X_T: Mapping from the original space to the target space at a computational temperature T.
- w_{ij} : Weights
- d_{ij} : Pairwise Distance in the target space.
- δ_{ij} : Pairwise Distance in the original space.

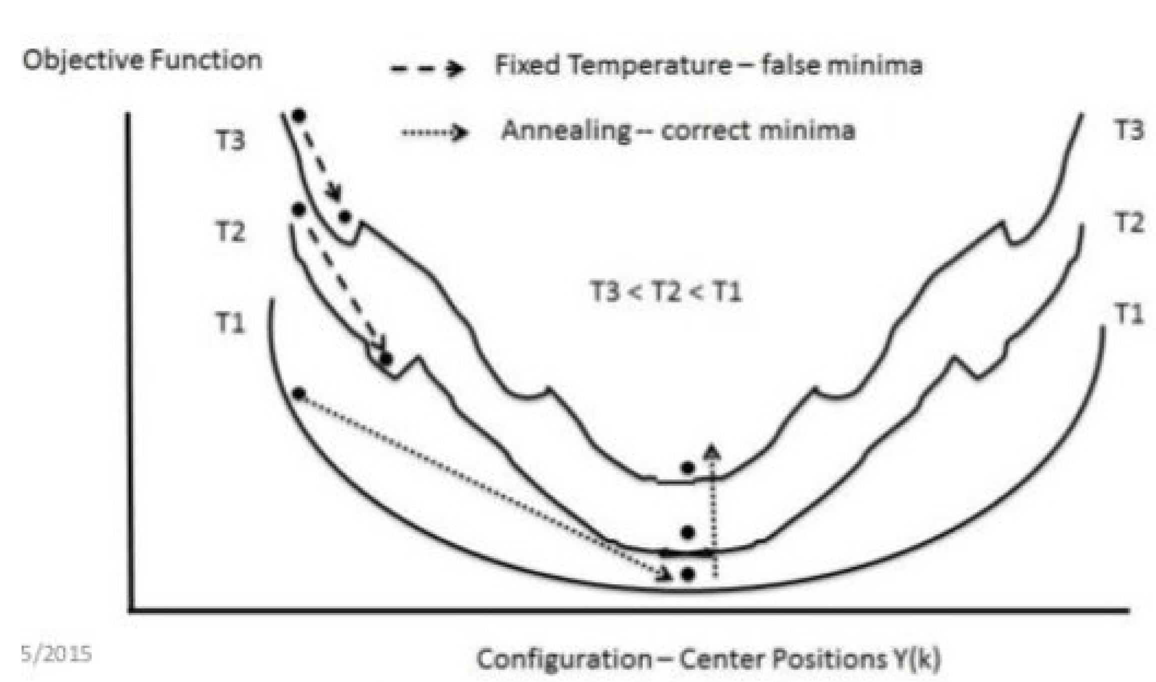

Figure 3: Deterministic Annealing
Source: DACIDR[?]

Results

- As the computational temperature decreases, the sparsity in the dataset increases.
- For a dataset of 22712 points, the average error in the calculation of stress is 0.0006.
- Even less number of far points approximate to the exact summation, thus the contribution of far points remains same.

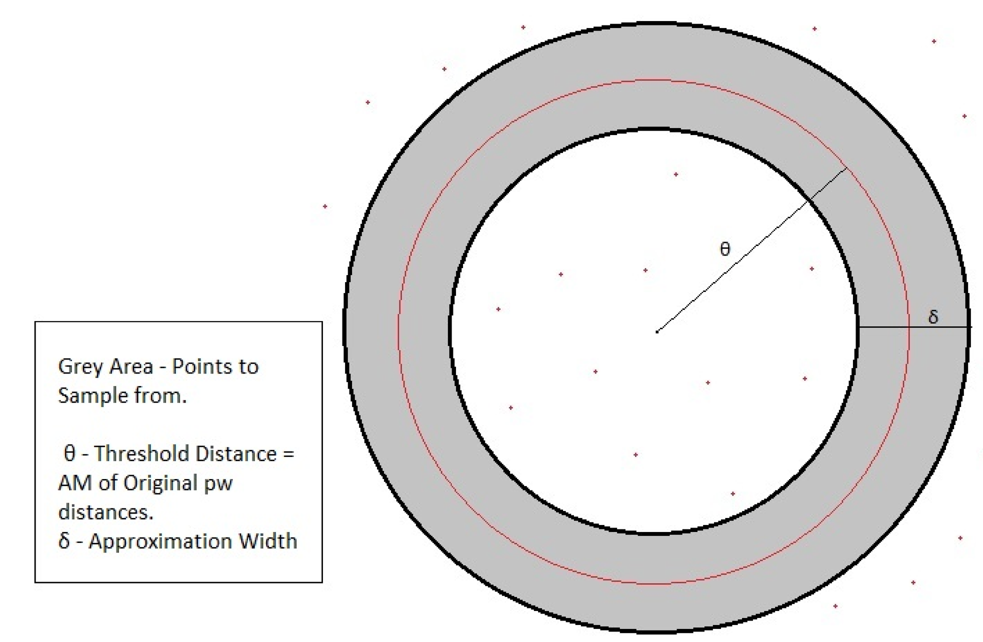

Figure 2: Near and Far Shells
Source: Created with MS-Paint

Table 1: Near and Far points

Temp.	Near	Far
0.19269	8062040	720
0.17391	7833665	229095
0.149102	2165944	5896816
0.10960	1884875	7874273
0.098917	25601	8037159
0.062343	3579	8059181
0.000000	451	8062309

Future Work

Instead of using two groups we can construct k groups/bins and the number of samples picked from each group varies accordingly with the distance from the concerned point. Also instead of randomly sampling from the original distribution, we can consider importance sampling wherein we sample from some other distribution using heuristics as in this case the contribution of points in the summation as a function of distance.

Acknowledgements

I am grateful to Dr. Geoffrey Fox for providing me this opportunity to work under him, and also for the continued guidance and support throughout the project. I would also like to thank Mr. Saliya Ekanayake for assisting me with the code and his help whenever I was stuck.

References

[BH86] Josh Barnes and Piet Hut. A hierarchical o (n log n) force-calculation algorithm. 1986.

[GM00] Alexander G Gray and Andrew W Moore. N-body'problems in statistical learning. In *NIPS*, volume 4, pages 521–527. Citeseer, 2000.

[GR87] Leslie Greengard and Vladimir Rokhlin. A fast algorithm for particle simulations. *Journal of computational physics*, 73(2):325–348, 1987.

[WBMB] Bo Xiao William B. March and George Biros. Approximate skeletonization kernel-independent treecode in high dimensions. *SIAM Journal on Scientific Computing.*

[YR] M. Rho H. Tang S.-H. Bae et al Y. Ruan, S. Ekanayake. Dacidr: deterministic annealed clustering with interpolative dimension reduction using a large collection of 16s rrna sequences. *ACM Conference on Bioinformatics, Computational Biology and Biomedicine.*