KeyBin2: Distributed Clustering for Scalable and In-situ Analysis

Xinyu Chen[1], **Matt Peterson**[1], **Jeremy Benson**[1], **Michela Taufer**[2], **Trilce Estrada**[1]

1. University of New Mexico, USA
2. University of Tennessee, Knoxville, USA

Correspondence: xychen@cs.unm.edu

Problem

Clustering high dimensional data is difficult because some features are correlated or noisy; distances are less meaningful; computation becomes expensive. Our previous work KeyBin[4] is a clustering method that uses keys and bins to learn from distributed data in parallel. It builds final high dimensional clusters from every 1-dimensional primary clusters. But it has the following limitations:

- Orthogonality assumption
- Overlapping problems
- Partitioning heuristic

Methods

KeyBin2 uses random projections and discrete optimizations to overcome the above limitations. Basic **methods** are:

- Random projection: a linear transformation that multiplies a random matrix to high dimensional data points to project them to lower dimensional spaces.
- Moving average smoothing: a approximation that captures the major characteristics.
- Calinski-Harabaz index: evaluates separation of unlabeled data. Higher is better.
- Model searching: uses Calinski-Harabaz index to score projections and choose the best.

Random Projections

Create transformation matrix, a_i are unit random vectors: $A_{(N\times N_{rp})} = [a_1\ a_2\ a_3\ \dots\ a_{N_{rp}}]$. A vector $x = (x_1, x_2, .., x_N)$ is projected to $x' = (x'_1, x'_2, .., x'_{N_{rp}})$ with $N_{rp} << N$[1].

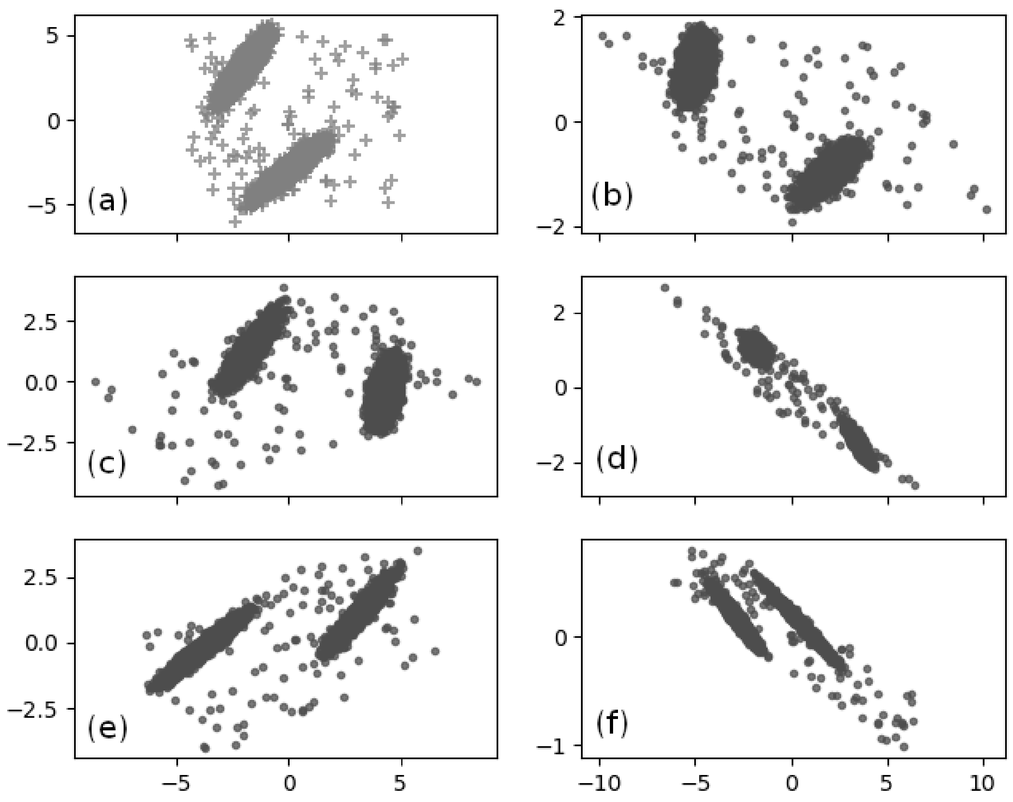

Figure 1: subfigure(a): 2 clusters with correlated features . subfigure(b)-(f) 5 projected spaces after random projection. (b) and (c) decorrelate and provide better separations. (d)(e)(f) do not provide good separations.

Partitioning Smoothed Histograms

In a 2-dimensional example, the algorithm builds histograms on x and y respectively. Because the smoothed curves over histograms are differentiable, we are able to find partitioning locations with 1^{st} and 2^{nd} derivatives.

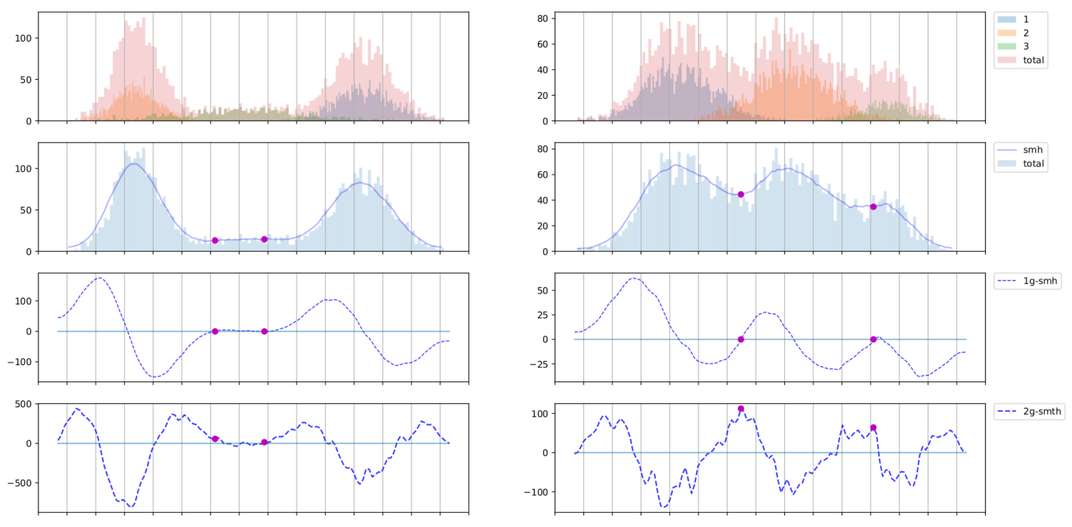

Figure 2: Row_1: the original histogram. Row_2: smoothed curve over the histograms. $Rows_3$: the first derivatives. $Rows_4$: the second derivatives. Pink dots: the partitioning locations.

Assessing Projected Subspaces

Due to the randomness, some projections provide better separation than others. We use a modified Calinski index[2] to assess each model and choose the best model.

$$cal = \left[\frac{B_Q}{W_Q}\right] \times \left[\frac{|Bins| - |Q|}{|Q| - 1}\right] \times log_2(|Q| - 1) \quad (1)$$

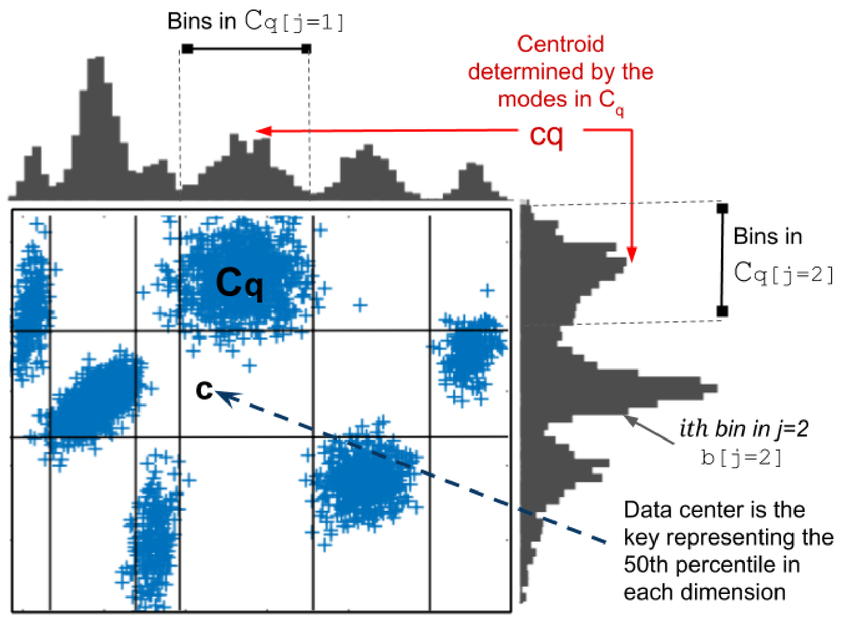

Figure 3: Illustration of the modified Calinski index for a 2-dimensional example.

Experiments and Results

KeyBin2 achieves f1scores 0.867 in clustering 1.28 million synthetic high dimensional data points. This accuracy higher than other clustering algorithms such as K-means++(0.821), parallel K-means($.695$) and pdsdbscan(0.445).

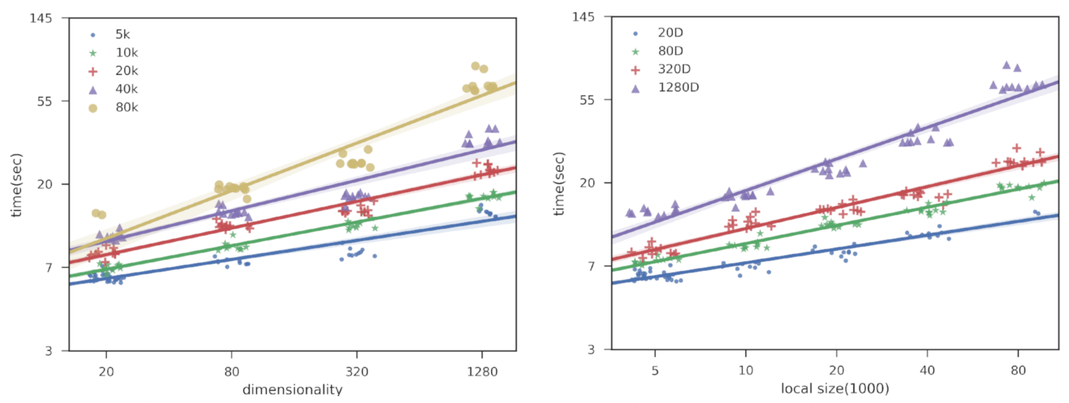

Figure 4: left:scale with number of dimensions. right:scale with number of points.

We use KeyBin2 to find clusters of stable status for protein folding trajectory data[3]. Align the clustering results with probabilities computed from coordinates. Some results(fig5left) are promising. Some clusters of transition phases(fig5right) need further study.

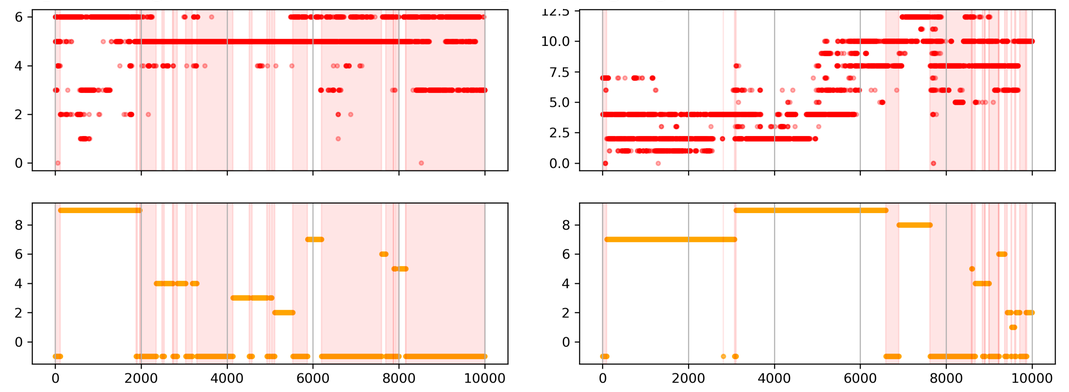

Figure 5: left:align transition phases of 1a0b-1 protein. right:align transition phases of 1a70-1 protein.

Conclusion

In this poster, we presented KeyBin2, the improved distributed clustering algorithm for scalable and in-situ analysis. KeyBin2 uses multiple random projections and modified Calinski-Harabaz index to overcome the orthogonal assumption and overlapping limitations of our previous work. It uses smoothing technique to get more robust partitioning locations. It is able to find clusters in correlated high dimensional spaces. We achieves good scalability and accuracy in synthetic data and real data.

Acknowledgments

The work is support by NSF grants *CAREER: Enabling Distributed and In-Situ Analysis for Multidimensional Structured Data* (NSF ACI-1453430) and *BIGDATA: IA: Collaborative Research: In Situ Data Analytics for Next Generation Molecular Dynamics Workflows* (NSF 1741057).
The authors would like to thank the UNM Center for Advanced Research Computing(CARC) for computational resources used in this work.

References

[1] Bingham, E., Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 245250. https://doi.org/10.1145/502512.502546.

[2] Caliski, T., Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-Theory and Methods, 3(1), 127..

[3] Zhang, B., Estrada, T., Cicotti, P., Taufer, M. (2014). Enabling in-situ data analysis for large protein folding trajectory datasets. In IEEE International Parallel and Distributed Processing Symposium.

[4] Chen, X., Benson, J., Estrada, T. (2017). keybin: Key-Based Binning for Distributed Clustering. In 2017 IEEE International Conference on Cluster Computing (CLUSTER) (pp. 572581). https://doi.org/10.1109/CLUSTER.2017.96