KeyBin2: Distributed Clustering for Scalable and In-Situ Analysis

Xinyu Chen
University of New Mexico
Albuquerque, New Mexico
xychen@cs.unm.edu

Matt Peterson
University of New Mexico
mpeterson@unm.edu

Jeremy Benson
University of New Mexico
jeremybenson@cs.unm.edu

Michela Taufer
University of Tennessee
taufer@utk.edu

Trilce Estrada
University of New Mexico
Albuquerque, New Mexico
estrada@cs.unm.edu

ABSTRACT

We present KeyBin2, a key-based clustering method that is able to learn from distributed data in parallel. KeyBin2 uses random projections and discrete optimizations to efficiently clustering very high dimensional data. Because it is based on keys computed independently per dimension and per data point, KeyBin2 can scale linearly. We perform accuracy and scalability tests to evaluate our algorithm's performance using synthetic and real datasets. The experiments show that KeyBin2 outperforms other parallel clustering methods for problems with increased complexity. Finally, we present an application of KeyBin2 for in-situ clustering of protein folding trajectories.

CCS CONCEPTS

• **Mathematics of computing** → **Cluster analysis**; **Exploratory data analysis**; • **Computing methodologies** → *Massively parallel algorithms*;

KEYWORDS

Scalable Clustering,Privacy Preserving,Big Data,Random Projection

ACM Reference format:
Xinyu Chen, Matt Peterson, Jeremy Benson, Michela Taufer, and Trilce Estrada. 2018. KeyBin2: Distributed Clustering for Scalable and In-Situ Analysis. In *Proceedings of 47th International Conference on Parallel Processing, Eugene, OR, USA, August 13–16, 2018 (ICPP 2018),* 2 pages.
https://doi.org/10.1145/3225058.3225149

1 INTRODUCTION

Clustering high dimensional data is difficult due to the so called "curse of dimensionality". Meanwhile scientific simulations are generating huge amount of such high dimensional data. Climate simulations and high-energy physics simulations [3] produce Terabytes or Petabytes data per day. Besides, privacy concerns makes the learning more challenging [4]. Medical or financial data are not large, but they are not allowed to transfer. We want to address this three-fold challenge of clustering high dimensional data in a distributed manner when they are not moved to a centralized location.

For many learning tasks, reducing data dimensionality can accelerate the converge speed and often make the underlying structures more obvious. Principle components analysis and random projection [1] provide good approximation of the distance between data points in the projected lower dimensional spaces. However, when the data are produced and stored on distributed locations, if they cannot be gathered to a centralized location, these two reduction methods may produce less effective approximations.

The goal of our research is to use dimensionality reduction techniques in learning unlabeled data when the full original data are bound to distributed locations. Our previous works [2] propose a clustering algorithm that works on partial ordering and histograms of data points. This algorithm learns from the data densities to avoid pair-wise distance computations. The proposed clustering algorithm reduces data movement overheads and protect individual data points from being reproduced away from their native locations. To get better accuracy, it is essential to collapse noisy features and reduce data to lower dimensional spaces. In our current research, we further improve the algorithm by projecting the data to much lower dimensional spaces compared to just collapsing the noisy features.

2 THE METHOD

We first use random projection to reduce the dimensionality dramatically. Fig. 1 illustrate the effect of random projection. Then a histogram is built upon each projected dimension. To avoid introducing a specific threshold parameter for building preliminary clusters, we apply smoothing techniques on the histograms and use the 2rd derivatives of the smoothed curves to find local minima as partitioning points. Fig 2 illustrates the partitioning method. Due to the random nature of random projection technique, we modify the Calinski-Harabasz index and compute the index from data densities to evaluate and choose the model with better clustering dispersion.

ICPP 2018, August 13–16, 2018, Eugene, OR, USA

ACM ISBN 978-1-4503-6510-9/18/08…$15.00
https://doi.org/10.1145/3225058.3225149

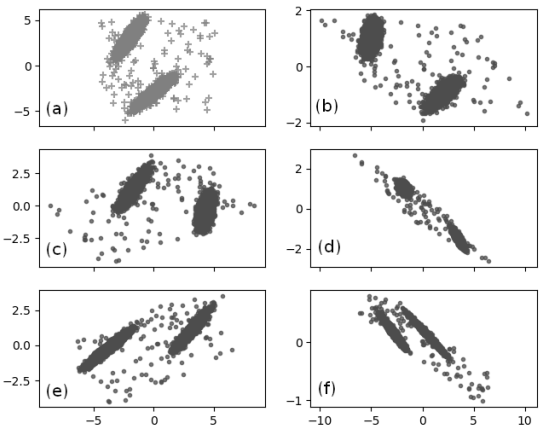

Figure 1: (a) original 2D data points. (b)-(f) projected points in space.

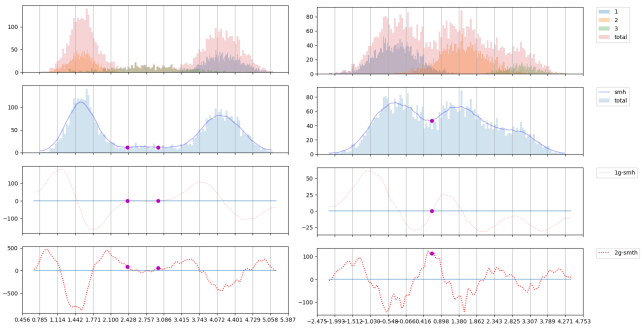

Figure 2: row1:original histogram showing 3 groups. row2:smoothed curve over histogram. row3:1st derivatives of smoothed histogram. row4: 2nd derivatives of smoothed histogram. Pink dots: partitioning points.

3 EXPERIMENT RESULTS

Experiments on both synthetic and real data have shown improved results in scalability and accuracy.

3.1 Tests on synthetic data

We run the *keybin2* algorithm with high dimensional synthetic data on distributed sites(up to 16 MPI processes). The dimensionality of data points increases from 20 to 1280. The number of data points increases from 5000 to 80000 per site(up to 1.2 million in total). Fig 3 shows our algorithm scales well in both cases.

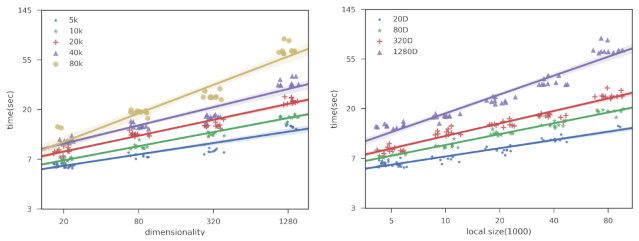

Figure 3: left:scale with number of dimensions. right:scale with number of points.

3.2 In-situ analysis on protein trajectories

We use KeyBin2 to find clusters of stable status for protein folding trajectory data [5]. Align the clustering results with probabilities computed from coordinates. Some results(fig 4left) are promising. Some clusters of transition phases(fig 4right) need further study.

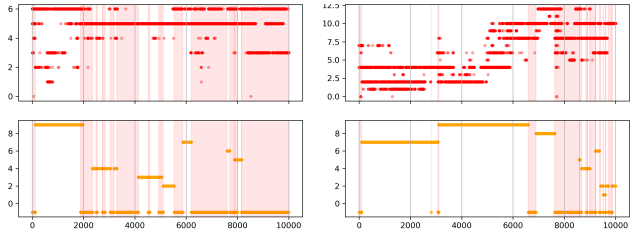

Figure 4: left:align transition phases of 1a0b-1 protein. right:align transition phases of 1a70-1 protein.

4 CONCLUSIONS

In this poster we presented the improved binning clustering algorithm *KeyBin2*. This parallel clustering algorithm uses bootstrapping and random projection methods to overcome the limitation of orthogonality assumption of our previous method (KeyBin). The rotation effect of random projection helps to separate overlapping clusters which are not solved in KeyBin. In this version, we eliminate a density threshold in the partitioning heuristics, thus producing more robust clustering results. With these improvements, KeyBin2 improves scalability and can deal with more complex data than its predecessor. Experiments show that our algorithm scales linearly when the number of data points or the dimensionality increases. Finally, we show the applicability of KeyBin2 for in-situ analysis of folding trajectories.

ACKNOWLEDGMENTS

The authors would like to thank the Data Science group at the University of New Mexico. The works is supported by NSF grant entitled CAREER: Enabling Distributed and In-Situ Analysis for Multidimensional Structured Data (NSF ACI-1453430). We also thank the UNM Center for Advanced Research Computing for computational resources used in this work.

REFERENCES

[1] Ella Bingham and Heikki Mannila. 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* (2001), 245–250. https://doi.org/10.1145/502512.502546

[2] Xinyu Chen, Jeremy Benson, and Trilce Estrada. 2017. keybin: Key-Based Binning for Distributed Clustering. In *Cluster Computing (CLUSTER), 2017 IEEE International Conference on*. IEEE, 572–581.

[3] Daniel A Reed and Jack Dongarra. 2015. Exascale computing and big data. *Commun. ACM* 58, 7 (2015), 56–68.

[4] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. 2014. Data mining with big data. *IEEE transactions on knowledge and data engineering* 26, 1 (2014), 97–107.

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