Scalable Big Data Clustering by Random Projection Hashing

Sayantan Dey, Lee Carraher, Anindya Moitra, and Philip A. Wilsey (PI)
Dept of Electrical Engineering and Computer Science, University of Cincinnati

UNIVERSITY OF Cincinnati

Abstract:
This project is developing a novel algorithm, called *Random Projection Hash* or RPHash. RPHash utilizes aspects of random projection, locality sensitive hashing (LSH), and count-min sketch to achieve computational scalability and linear achievable gains from parallel speed up. The approach is data agnostic, minimizes communication overhead, and has a priori predictable computational time. The system is deployable on commercially available cloud resources running the Spark implementation of MapReduce. The RPHash solution will have a wide applicability to a variety of standard clustering applications while this project will focus on a subset of clustering problems in the biological data analysis space. RPHash also combats de-anonymization attacks inherently resulting from its algorithmic requirements thus addressing requirements involving the handling and privacy protection of health care data as well as the inherent privacy concerns of using cloud based services. Furthermore, RPHash will allow researchers to scale their clustering problems without the need for specialized equipment or computing resources. The proposed cloud processing solution will allow researchers to arbitrarily scale their processing needs using virtually limitless commercial processing resources.

RPHash Algorithm:

1. Project high-D → low-D
2. LSH low-D vectors
3. Count-min sketch
4. Top k counters are centroids

Clustering Accuracy:

HAR Data	ARI	Time (sec)
KMeans	0.461	24.746
RPHash	0.363	0.4838
RPHash-Dis	0.48	8.116

Scalability (time as dimensions increase):

Dimension	KMeans	RPHash	RPHash-Dist
100	3.9656	0.4890	7.54
500	32.92	0.5594	7.87
1000	142.2341	0.7206	8.51
1500	237.0007	0.8233	9.48
2000	366.0743	0.8796	10.73
2500	431.5876	0.9997	11.43
3000	542.0223	1.1122	13.39
3500	631.8423	1.2421	12.97
4000	741.915	1.3157	13.78
4500	811.3911	1.3986	14.40
5000	909.223	1.5095	15.14
5500	975.2703	1.5977	15.98
6000	1076.882	1.6805	17.24
6500	1187.6062	1.7933	18.02

Contact:
Philip A. Wilsey (PI): wilseypa@gmail.com
Sayantan Dey: deysn@mail.uc.edu
Lee Carraher: leecarraher@gmail.com
Anindya Moitra: moitraaa@mail.uc.edu

RPHash in a map-reduce deployment