Abstract

RPHash is a streaming algorithm for data clustering based on frequent itemset counting of locality sensitive hash (LSH) collisions generated by multiple randomly projected data vectors. This proposed algorithm called Random Projection Hashing or RPHash trades sub-optimal computationally performance for improved memory efficiency over other clustering algorithms. Memory efficiency, a premium in streaming and distributed algorithms allows RPHash to be run on a variety of streaming data environments. Two additional features, temporal weighted and privacy are also provided automatically by RPHash. By performing multiple projections stilling recent advances in the fast johnson lindenstrauss transform (FJLT) we are able to attain a temporally weighted, streaming clustering algorithm with log-linear complexity growth and intrinsic data security.

RPHash Algorithm

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Locality Sensitive Hashing</strong></td>
<td><strong>Multi-Prob Random Projection</strong></td>
</tr>
<tr>
<td><strong>Fast Johnson-Lindenstrauss Transform</strong></td>
<td><strong>Projected KMeans</strong></td>
</tr>
<tr>
<td><strong>Locally Sensitive Hashing</strong></td>
<td><strong>Projected KMeans</strong></td>
</tr>
<tr>
<td><strong>Universally Generative Lattices</strong></td>
<td><strong>Lossy Frequency Counting</strong></td>
</tr>
<tr>
<td><strong>Low Inter-Process Communication</strong></td>
<td><strong>Data Security</strong></td>
</tr>
<tr>
<td><strong>Clustering on a Grid</strong></td>
<td><strong>Cloud Based Resource Allocation</strong></td>
</tr>
</tbody>
</table>

Key Components

- Multi-Prob Random Projection
- Fast Johnson-Lindenstrauss Transform
- Locality Sensitive Hashing
- Universally Generative Lattices
- Lossy Frequency Counting

Applications

- Bioinformatics
- Medical Imaging
- Computer Vision
- Recommender Systems
- Social Network Analysis
- Market research

Early Results

- Time Vs Inter-Cluster Variance
- Precision-Recall vs Inter-Cluster Variance

Security Via Random Projection

- Improved Data Security and HIPAA Compliance Concerns made Commercial
- Cloud Computing an Economic Resource
- Random Projection - RPHash can provide significant Resistance to
- Denial of service - a side Random Projection is very sensitive to this.

Contact

- Open source available at: github.com/wilseypa/rphash
- Philip A. Wilsey (PI): wilseypa@gmail.com
- Lee Carraher: leecarraher@gmail.com
- Anindya Moitra: moitraaa@mail.uc.edu