Agglomerative mean shift clustering embedding query set compression

Damre Suraj S[1].,L.M.R.J.Lobo

Department of Computer Science and Engineering

Walchand institute of technology Solapur , Maharashtra ,India
Email- surajdamre@gmail.com

Abstract-

The agglomerative mean shift clustering is known as non parametric clustering because it does not require any prior knowledge of number of clusters and it does not constrain about their shape. This approach provides a better quality than other clustering approach .In this algorithm constructing a family of d-dimensional hyper ellipsoids to cover the query set and the point inside each hyper ellipsoid will converge to local maximum. After that we use centres to form new query set which is the compressor of original to reduce the time complexity. So at each iteration level clustering is done. This is fast and time complexity of this less because of the compression of query set.

1.INTRODUCTION

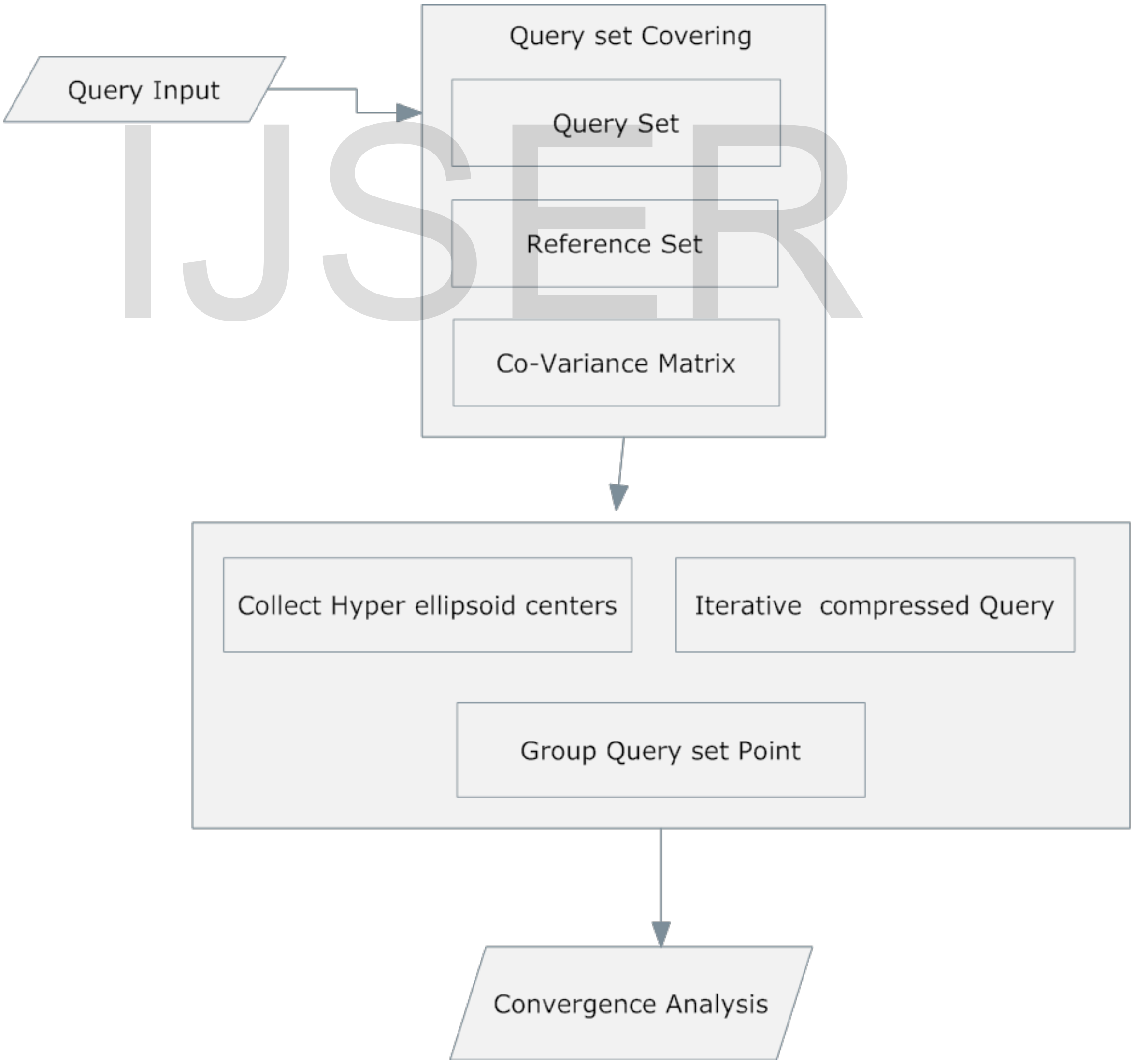
Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendrogram. This tree graphically displays the merging process and the intermediate clusters.

Agglomerative mean shift algorithm is to construct a family of d-dimensional hyper ellipsoids to cover the current query set Q. The points inside each hyper ellipsoid will converge to a common local maximum KDE via Mean shift. Then we use centres of these hyper ellipsoids to form a new Query set as the compressor of original one. We run this iteration several times until the convergence is done. At each iteration level clustering is done by grouping the current Query points according to the hyper ellipsoids.

The steps of iteration are

- Create a binary cluster tree with single article clusters as its leaf nodes and a root node containing all the articles.
- Maximize the Kernel Density Estimator (KDE).
- Perform nonparametric mode-seeking and clustering.
- Use this as grouping the points in a query set Q according to the modes they converge on a reference set R.
- Generate an efficient query set compression mechanism to accelerate MS-clustering.

2.System architecture



3.MODULES

1) Dataset Collection & Input Query

2)Hyper ellipsoid Computation

3) Sample Query set Creation

4) Query Set Compression

5) Convergence Analysis

MODULES DESCRIPTION:-

- Dataset Collection & Input Query

  The document dataset is collected.

  The stop words from the collected documents are removed.

  This is called preprocessing

  The input Query is given for the retrieving process.

  The query is preprocessed before proceeding with further analysis.

- Hyper ellipsoid Computation

  – The initial thing in finding the hyper ellipsoid is to find the Mahalanobis distance between the query and the document along with the co-efficient matrix.

  – The basic idea is to construct a family of d-dimensional hyper ellipsoids to cover the current query set Q.

  – For a given query point, the hyper ellipsoid is constructed from a lower QB function of Kernel Density Estimation.

- Sample Query set Creation

  – After forming number of hyper ellipsoids, empirically, it is observed that the number of the covering hyper ellipsoids is much smaller than the size of Q.

  - Then take the centers of these hyper ellipsoids to form a new query set with size dramatically reduced.
  - Such a query set covering procedure can be iteratively run until a sample set of hyper ellipsoids are obtained.

- Query Set Compression
  - Given the currently constructed hyper ellipsoid set S_0, may take the centers of the hyper ellipsoids in it to form a compressed query set.
  - The above presented set covering operation can be directly applied on Q1.
  - After sufficient iterates until convergence, we will obtain a sparse enough query set Q_∞.
  - At each iteration level l, we could group points in Ql according to their associated hyper ellipsoids in S_l.
  - Such a query set compression framework naturally leads to an agglomerative clustering of Q_0.

- Convergence Analysis
  - In this module, the convergence property of query set size sequence that is generated using the above mentioned algorithm is analyzed.

4. Conclusion

In this proposed work, an Agglomerative mean shift algorithm is developed to accelerate the widely applied Mean-Shift Clustering method. This algorithm provides an efficient hyper ellipsoid query set covering mechanism that reduces the MS iterations during Clustering. It seems to be more flexible than any other existing acceleration algorithms such as IFGT-MS and LSH-MS.

Agglomerative mean shift algorithm is a fast, stable, and accurate Mean shift clustering algorithm that can achieve competitive solutions.

5. REFERENCE

1) M. Carreira-Perpinan, “Acceleration Strategies for Gaussian Mean-Shift Image Segmentation,” Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 1160-1167, 2006.

2) Y. Cheng, “Mean Shift, Mode Seeking, and Clustering,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 7, pp. 790-799, July 1995.

3) Y. Zho and G. Karypis, "Hierarchical Clustering Algorithms for Document Datasets," Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 141-168,

Mar. 2005.

4) M. Allain, J. Idier, and Y. Goussard, "On Global and Local Convergence of

Half-quadratic Algorithms," IEEE Trans. Image Processing, vol. 15, no. 5, pp.

1130-1142, May 2006.

5) M. Carreira-Perpinan."Fast nonparametric clustering with Gaussian blurring mean-shift". In International Conference on Machine Learning, pages 153–160, 2006.

6) M. Carreira-Perpinan."Gaussian mean-shift is an em algorithm".IEEE TPAMI, 29(5):767–776, 2007.

7) Y. Cheng. "Mean shift, mode seeking, and clustering".IEEE TPAMI, 17(7):790–799,

1995.

8) D. Comaniciu and P. Meer. "Mean shift: A robust approach toward feature space analysis". IEEE TPAMI, 24(5):603–619, May 2002.

9) I. Davidson and S. Ravi. "Clustering with constraints: Feasibility issues and the k-means algorithm". In International Conference on Data Mining. SIAM, 2005.

10) I. Davidson and S. Ravi. "Hierarchical clustering with constraints: theory and practice". In principles and Practice of Knowledge Discovery in Databases (PKDD), 2005.