Peer Review
Reviewing Paper : Kernel k-means, Spectral Clustering and Normalized Cuts

Aditya Patil Bhasker Goel Kartikeya Saxena and V Anirudh

Department of Computer Science, Indian Institute of Technology, Guwahati
*Corresponding author: {aditya18, bhasker18, kartikey18a, anirudh18}@iitg.ac.in

1 Summary

- The base static algorithm is a combination of spectral clustering and kernel k-means. The algorithm takes k, the number of clusters we want to find, and k' the number of nearest neighbours used to construct a graph for spectral clustering as parameters. As Spectral Clustering is primarily a graph clustering algorithm, a k'-nearest neighbour graph is constructed from the points in the dataset. After this graph is obtained, an iteration of spectral clustering is run. Then, the eigen-vectors of the corresponding laplacian matrix obtained from spectral clustering are used to provide the initial clustering on the database. Then kernel k-means is performed on this initial clustering to obtain better results. Using a kernel method equips the algorithm to deal with non-linear datasets. This combination of both the clustering algorithms produces mentionable results.
- The proposed incremental algorithm handles addition, deletion on non-linear data sets along with detection of duplicate points. Initially clustering obtained from static version is stored in a csv file for future use. Then, the optimal number of operations that can be performed and representative points for each cluster is calculated. The Eigen Gap Heuristic is employed to determine the number of clusters to be formed. For each iteration of addition, first the data is checked for duplicate point and then the inserted point is added to the cluster with nearest representative point. For deletion, the point is simply marked deleted. After the number of additions and deletions reach the optimal number of operations, static algorithm is re-run on the complete data.
- Incremental algorithm achieves a significant reduction in time complexity. Static algorithm takes $O(n^3)$ time where n is the number of points. In static algorithm calculation of representative points takes $O(n)$ time and after this for the next optimal number of operations time taken is $O(q^{3/2})$ where q is the number of representative points, also $q << n$. But these improvements come at the cost of accuracy, strict accuracy of test data sets is below 70%.

2 Key Strengths of the Project

- The incremental algorithms is seen to perform many times better than the static algorithm. Also the report presents these time, memory and accuracy metrics elaborately.
- Duplicate Detection is dealt with, which is something that is very easy to ignore; and yet might save time in some situations.
- The calculation of representative points in reducing the computation cost and time.

3 Key Weaknesses of the Project

- Strict accuracy is observed to be low (around 50-60%) and the concept of relaxed accuracy has been implemented incorrectly. (The team's implementation will always give the relaxed accuracy to be 100%). Instead, when a set of clusters in Clustering 2 is mapped to a particular cluster in Clustering 2, then some cluster belonging to this set should not be mapped to yet another cluster in Clustering 1. Concretely speaking, the concept should implement a One-to-Many mapping between the clusterings while the team has implemented Many-to-Many mapping leading to the relaxed Accuracy always being 100%.
- The datasets used were extremely small in size so not much insights could be gained on the performance of the suggested incremental algorithm. Also, major results have been presented only on toy datasets while very less emphasis is given to real life datasets.

- The suggested algorithm runs the static algorithm again after a certain number of insertions or deletions, which will prevent the incremental algorithm to provide speedups with increasing data.
- Deletion of points was not handled step by step, instead the deletion was handled when static algorithm ran.
- The incremental algorithm consumes significantly high memory when compared to the static run, which may cause problems in warehousing environments.

4 Suggested Improvements in the Project

- Compared to the speedup, loss in accuracy is too much. This trade-off could be made more balanced by reducing the number of optimal points in every iteration.
- Points are deleted lazily. The team may come up with some method to deal with detection immediately.
- The accuracy measure for relaxed accuracy does many to many mapping of clusters. Instead one to many mapping of clusters should have been done.
- Small test datasets may not provide accurate measure of the performance of the algorithm. Thus, larger datasets should be included for testing.

5 Suggested Improvements in the Report

- Results of both the algorithms could be compared and visualized properly with the help of graphs.
- Latex symbols are not properly used and some symbols are missing. Could use better maths modules.
- Tables are scattered all over the report. Could use one or two tables to convey the results in a concise manner.
- Minor mistakes in flow charts, grammar and mathematical symbols could be corrected.
- Detailed explanation could be provided for the formulae used in the algorithm.

6 Overall rating for the report on the scale of 0 to 10

All things considered the report was good. We think the overall rating for the report should be **8.5**.
Link to report: Group 6