Wan D. Bae - Data Science (ITE4005)
컴퓨터전공 2013012289 한기훈
May 24th, 2017

Report for Assignment3 - Clustering

1. Summary of algorithm.

The clustering algorithm I implemented is based on DBSCAN algorithm.

It follows the following procedure.

- Read object data from input file.
- Generate cluster list.
  - Select one unvisited object from data.
  - Check if noise or not.
  - If not noise, expand cluster as many as possible.
- Write into output file

2. Detailed description of code.

This code has three major parts.

A. `readDataFromInput()`

Read object data from given path. Since each line contains only three numbers, I don't have to consider newline character and parsing methods.

In this function, I define

```
map<int, pair<bool, pair<double, double> > > objects
```

and store each data from each line with this code

```
objects.insert(make_pair(obj_id, make_pair(false, make_pair(x, y))));
```

This function return variable `objects` that defined in previous line.

B. `DBSCAN()`

This function has some internal function implemented with lambda function.

- `getDist()`

It get two objects and calculate distance between two objects, and return it.

- `regionQuery()`

It get one object via argument and find all other objects that is neighbor of it.

Store object_id of neighbor objects into `set<int> N` (except object from argument);

Return variable `N` that stores object_id of each neighbors.

- `expandCluster()`

Expand cluster using below method.

* Find neighbors of each object that is not visited yet, and if size meets condition (size >= MinPts) put it into clusters. Iteratively doing this process until no more neighbors group is found.

Procedure starts from line number 133 in main.cpp

Iteratively get each object which is not visited yet.

Set selected object as visited, and get neighbors list.

If size of neighbors list is equal or bigger than MinPts, call `expandCluster()` function.

After clustering process, sort each cluster by size of cluster in descending order.

Then leave only `num`(number of clusters: given from main function args) and erase everything.

Return remain cluster list.

C. `printResult()`

This function get variable `cluster_list` with type `vector< set<int> >`.

Then, iterate each set in `cluster_list` and write result into output file.

This code specifies output file path.

```
string output_file_path = output_path + "_cluster_" + to_string(i) + ".txt";
```

Then, in each output file, just print one object_id in each line.

3. How to compile this code.

Just type 'make' in terminal. Or, please type below line.

```
$ g++ -O2 -o clustering.exe main.cpp --std=c++11
```

Above task will generate 'clustering.exe' file into same directory.

Compiler must support C++11 standard.

4. Any other specification.

Using sample data, test program generates following results (accuracy).

- input1.txt : 98.97037% accuracy
- input2.txt : 94.86598% accuracy
- input3.txt : 99.97736% accuracy