



Introduction

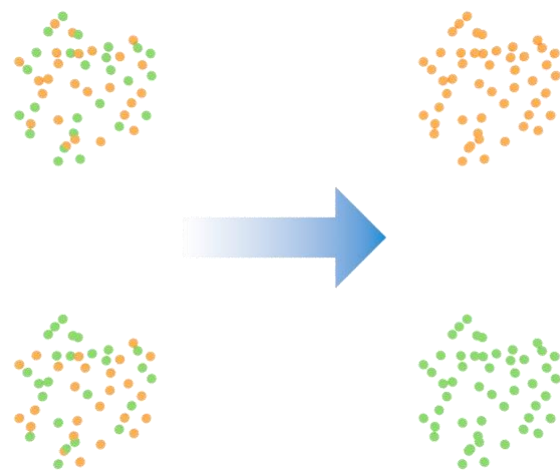
Data Clustering is defined as grouping together objects that share similar properties. These properties can be anything as long as it is possible to measure and compare them.

Areas where clustering is much used are:

- machine learning
- search engines
- image analysis

Below is an example of clustering on a set of 100 points. The red and green points suggest two different clusters. At the left side of the figure the points have been randomly distributed between two clusters. The results after clustering are shown to the right. Here we can see that the algorithm has found the structure of the clusters and correctly organized the data by putting the points in the correct groups.

Note that no points are physically moved, only what cluster they belong to have been changed.

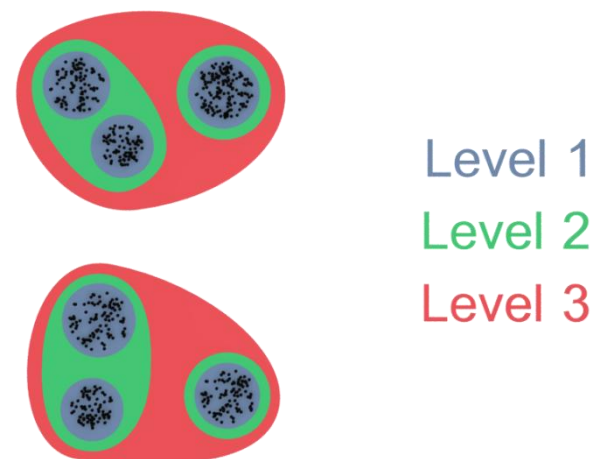


Clustering approaches

In clustering theory there are two main approaches; partitional and hierarchical.

In partitional clustering the goal is to partition the objects into a static number of clusters. The main advantage with partitional clustering is speed.

The goal in hierarchical clustering is to have a multi-layered solution. For example two groups where each of them are made up from smaller sub groups. The figure below is an example of this. Level 2 and level 3 have clusters which also contains sub-cluster from the levels below.



The algorithms

We have developed and tested several algorithms suitable for clustering large data sets. Two of these were deemed good enough and was tested in our experiments phase.

Our aim was to integrate the following characteristics:

- hierarchical clustering
- local search approach
- online clustering

In large datasets where the data are related, changes will almost always affect other data. A tough challenge was how to minimize the changes only to the affected area of the dataset. This is known as the “local search approach”

Experiments

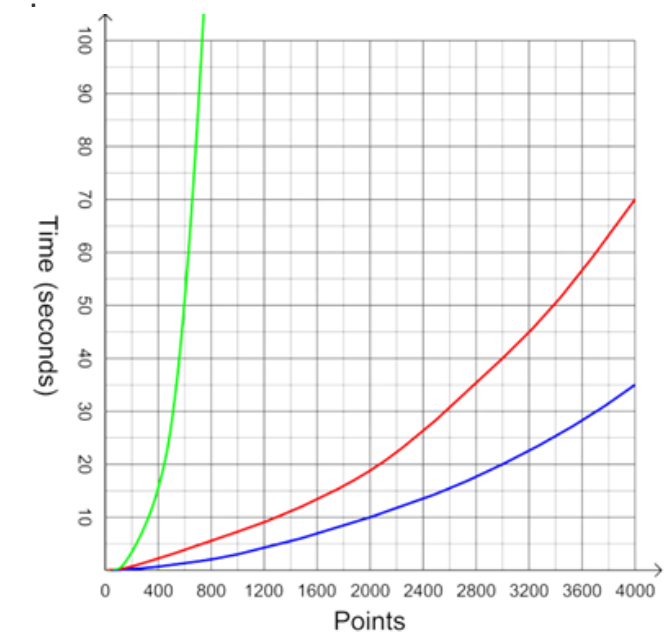
To evaluate the performance and quality of these algorithms, we have developed a software solution which enables us to run different algorithms and create a graph with time and quality as the axes. The software can be set to run many times on the same problem to generate an average result. We have also implemented some other clustering methods to compare against. Furthermore, we have tested the algorithms against a well known “state-of-the-art” algorithm. This method will almost always reach the best possible result, but scales very poorly.

Results

We have run the algorithms on several different problems, varying in both size and complexity. Our experiments tell us our algorithms are:

- Scalable
- Quicker than the “state-of-the-art” algorithm on large problems
- Able to find all sub clusters

The figure below shows how our algorithms scale compared to the “state of the art” algorithm (green line). Both the ‘cluster switch’ (red line) and the ‘cluster mover’ (blue line) clearly scales better.



Conclusion

The results from our experiments show that the local search approach used, increases scalability. Our algorithms perform high quality clustering within decent time.

Compared to the 'state of the art' algorithm, they provide, on large problems, a somewhat lower quality solution, but it requires less clustering-time. The difference in quality mostly represents internal structure in natural clusters, meaning the algorithms almost always finds natural clusters and sub-clusters before the “state-of-the-art” algorithm does.