Clustering algorithms on GPUs

Iterative clustering is a technique used in unsupervised machine learning to cluster similar data points together. It involves iteratively partitioning the data points into groups, based on their similarity, until a satisfactory clustering is achieved.

Input: set of vectors V, threshold τ,
number of centroids per iteration k
Output: set of clusters of vectors $\mathcal{C}$
$\mathcal{C} \leftarrow \emptyset$
while $|V| \geq k$ **do**
  $C_1 \leftarrow$ new $\{\}, \ldots, C_k \leftarrow$ new $\{\}$
    // initialize k new empty clusters
  select $c_1, \ldots, c_k \in V$ *// select k centroids*
  for $v \in V$ **do**
    $j \leftarrow \arg\min_{1 \leq i \leq k} \hat{d}_J(c_i, v)$
      // find closest centroid c_j
    if $\hat{d}_J(c_j, v) \leq \tau$ **then**
      // if c_j is close enough
      $V \leftarrow V \setminus \{v\}$ *// ... move v to C_j*
      $C_j \leftarrow C_j \cup \{v\}$
  $\mathcal{C} = \mathcal{C} \cup \{C_1, \ldots, C_k\}$ *// add new clusters to $\mathcal{C}$*
for $v \in V$ **do**
    // add remaining vectors as singleton clusters
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{\{v\}\}$
return $\mathcal{C}$

The iterative clustering process typically starts with an initial set of cluster centres, which can be randomly selected or obtained through some other means. Then, each data point is assigned to the nearest cluster centre based on some distance metric, such as Euclidean distance or cosine similarity (pseudocode gives an example).

In this project, we want to investigate the advantages and the algorithmic problems of using GPU parallelism to speeding up iterative clustering. Additionally, students are encouraged to explore estimators and heuristics for optimizing the selection of comparison similarity patterns and centroid choices.

We recommend developing the new implementations using the OpenACC. By doing so, students can give more attention to investigating algorithmic challenges rather than focusing on low-level implementation details.

Project Argument	Algorithms for Data Science		
TA / Support	Prof.Vella/Dott. Pichetti		
Team members	2		
Project type	Theory	Code	Tool
	60	30	10
Involved metrics	**Time/Similarity**		

Relevant papers:

https://ieeexplore.ieee.org/abstract/document/10027183