K-Means Clustering: comparison between a sequential and parallel implementation

Beatrice Paoli
beatrice.paoli@stud.unifi.it

Abstract

The goal of this paper is to present two implementations of the k-means clustering algorithm: a sequential version and a parallel one. Both are written in C++, with the parallel version obtained with the use of OpenMP. This paper also provides a performance analysis and comparison between the two versions in terms of execution times and speedup and how this values change with different inputs (dataset size and number of centers) and different numbers of threads.

1. The Algorithm

The K-means algorithm is a parametrized clustering technique that allows to partition a dataset of points or observations into K clusters. Points in the same cluster minimize the distance from the center of the cluster (called *prototype*). Many metrics of distances can be used depending on the dataset to cluster, for this implementation the dataset is comprised of 2D points and distances are measured with the Euclidean distance.

The algorithm is composed by few steps, with two main phases alternated between each other: the *Assignment* phase and the *Update* phase.

Algorithm 1 K-Means Clustering

Require: K = number of clusters to create

```
Select K points as initial centroids of the clusters
while Centroids keep changing do
    for each point p do
        Assign p to the cluster with the closest centroid
    end for
    for each cluster c do
        Update the centroid of c
    end for
end while
return clusters
```

The algorithm converges after few steps to a local minimum, depending on the choice of the initial centroids. In the implementation presented in this paper, the starting centroids are chosen randomly from the dataset in input. Therefor, different runs of the algorithm on the same dataset and the same number of centroids can yield different results. Other initialization methods can be used to obtain the gloabl optimum more consistently.

2. Implementation

The details of the implementations of the function *kMeansClustering()* are presented in the following paragraphs.

2.1. Classes

- **Point**: this class is used to represent the points of the input dataset to cluster. It has three members: x and y for the coordinates, and *clusterId* for the id of the cluster to which the point has been assigned. The class also has two constructors and the method *dist()* to compute the Euclidean distance between two points.
- **Cluster**: this class is used to represent the clusters created by the algorithm. Each cluster has an *id* ranging from 0 to $K-1$, a vector of objects of type *Point* containing the points assigned to the cluster, a *Point* for the current mean or centroid of the cluster, and two fields to compute the partial sums of all points for each coordinate, *tempSumX* and *tempSumY*. Aside from the constructor, the class has two main methods:
  - *addPoint()*: is used during the Assignment phase; it adds a point to the dataset and adds its coordinates to the partial sums.
  - *updateCentroid()*: is used during the Update phase; it computes the new mean of the cluster by using the sums of the coordinates computed before and the size of the list of points assigned. After the update, *tempSumX* and *tempSumY* are resetted to 0, ready for a new iteration of the algorithm.

2.2. kMeansClustering

3. Roba da togliere

3.1. Miscellaneous

Compare the following:
`$conf_a$` $conf_a$
`$\mathit{conf}_a$` conf_a
See The TEXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [3], and subsequently developed by Alpher and Fotheringham-Smythe [1], and Alpher *et al.* [2]."

This is incorrect: "... subsequently developed by Alpher *et al.* [1] ..." because reference [1] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.*

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [1, 3, 4] to [3, 1, 4].

4. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

4.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high.

4.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures **??** and 1. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

4.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

4.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [4]. Where appropriate, include the name(s) of editors of referenced books.

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

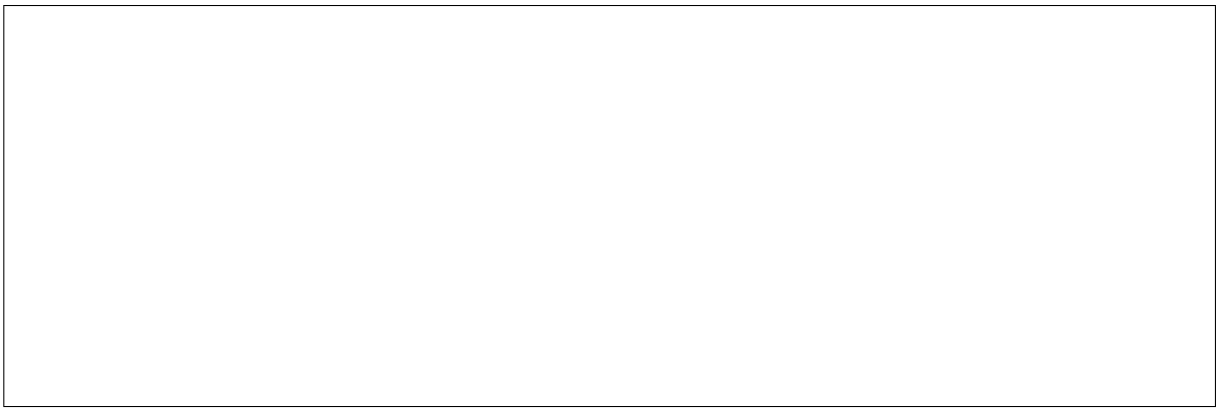

Figure 1. Example of a short caption, which should be centered.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

4.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

4.6. Color

Color is valuable, and will be visible to readers of the electronic copy. However ensure that, when printed on a monochrome printer, no important information is lost by the conversion to grayscale.

References

[1] A. Alpher, , and J. P. N. Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[2] A. Alpher, , J. P. N. Fotheringham-Smythe, and G. Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[3] A. Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[4] Authors. The frobnicatable foo filter, 2006. ECCV06 submission ID 324. Supplied as additional material `eccv06.pdf`.

5. Appendix

If your course project is part of a larger project from another class or research lab, please fill in this section and clearly spell out the following items:

1. Explicitly explain what the computer vision components are in this course project;
2. Explicitly list out all of your own contributions in this project in terms of:
   (a) ideas
   (b) formulations of algorithms
   (c) software and coding
   (d) designs of experiments
   (e) analysis of experiments
3. Verify and confirm that you (and your partner currently taking CS231A) are the sole author(s) of the writeup. Please provide papers, theses, or other documents related to this project so that we can compare with your own writeup.