An Improved K-Means Clustering Algorithm for One Dimensional Data

Ryan Froese
Dept. of Computer Science
University of Manitoba
Winnipeg, MB, Canada
froeser5@myumanitoba.ca

James Klassen
Dept. of Computer Science
University of Manitoba
Winnipeg, MB, Canada
klass167@myumanitoba.ca

Tyler Loewen
Dept. of Computer Science
University of Manitoba
Winnipeg, MB, Canada
loewent4@myumanitoba.ca

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Body

A. Algorithm Description

This algorithm needs a couple requirements in order to function properly. Firstly, there must be no duplicate cluster centroids. Secondly, every cluster must contain at least one item in it. Another thing to note is that if a datapoint has the same distance to two cluster centroids, our algorithm favours the cluster with the larger centroid (Although the chance of this happening is vanishingly small in real world datasets due to floating point precision. As such, it is ignored in this proof). We also assume there are at least two clusters ($k \geq 2$). In this proof we use the concept of cluster borders separating the data set into its clusters. This is an integral concept as the algorithm leans heavily on it.

To prove correctness of the algorithm, we must prove the following:

- Each item in the dataset will be assigned to the cluster with the closest centroid, and
- The cluster borders cannot cross each other (even if the algorithm is done in parallel), as this would cause sum/count calculations to be incorrect.

Setup

Algorithm inputs:

- $D = \{d_{11}, d_{12}, \ldots, d_{1n_1}, \ldots, d_{kn_k}\}$ where the dataset D contains elements d_{ij} where i is the cluster and j is the index within the cluster.
- k, the number of clusters
- $C = \{c_1, \ldots, c_k\}$ where C is a set containing the centroids of each cluster where each c is unique and $c_1 < c_2 < \cdots < c_k$. This sequence is constant until the very end of the algorithm.
- $S = \{s_1, \ldots, s_k\}$ where S is a set containing the sums of data points in each cluster
- $N = \{n_1, \ldots, n_k\}$ where N is a set containing the number of data points in each cluster
- $N_{\text{sum}} = n_1 + \cdots + n_k$ where N_{sum} is the total number of data points in D

In order to better visualize the intuition of the algorithm, we model the data and associated clusters as a sorted list of data points in clusters separated by cluster borders, i.e. the first border separates clusters 1 and 2, etc.. The set of all cluster borders positions is $B = b_1, \ldots, b_{k-1}$. Visually, this looks like the following diagram:

Outline of Proof of Property 1

Property 1.1: Data points can only be closest to one of the clusters they're adjacent to or inside of, i.e. d_{ij} is closest to either c_{i-1}, c_i or c_{i+1}

Definition 1.2: Definition: a cluster border b_i is in the correct location if all data points before bi are closer to c_i than c_{i+1}, and all data points after bi are closer to c_{i+1} than c_i.

Property 1.3: An equivalent statement: If a cluster border b_i is in the correct location, the item x immediately before b_i satisfies $x < (c_i + c_{i+1})/2$, and the item y immediately after b_i satisfies $y > (c_i + c_{i+1})/2$

Property 1.4: Once the first i cluster borders have been put in the correct location, it follows that all data points before b_i have been assigned to the correct cluster

Property 1.5: Once the last cluster border b_{k-1} has been put in the correct location, all data points after b_{k-1} have been correctly assigned to the last cluster.

Conclusion: By properties 1.4 and 1.5, once the algorithm terminates, all data points have been assigned to the correct cluster.

$$D = \underbrace{d_{11}, d_{12}, \ldots, d_{1x_1}}_{\substack{s_1 = d_{11} + \cdots + d_{1x_1} \\ c_1 = \frac{s_1}{x_1}}} \overbrace{\mid}^{b_1 = x_1} \underbrace{d_{21}, d_{22}, \ldots, d_{2x_2}}_{\substack{s_2 = d_{21} + \cdots + d_{2x_2} \\ c_2 = \frac{s_2}{x_2}}} \overbrace{\mid}^{b_2 = x_1 + x_2} \ldots \overbrace{\mid}^{b_{k-1} = x_1 + x_2 + \cdots + x_{k-1}} \underbrace{d_{k1}}_{s_k}$$