K-Means clustering variants and improvements

Stefan Sebastian 242

Overview

- K-Means classic algorithm
- Improvements: computation time, seed selection
- Variants: Bisecting K-means, Genetic Algorithms
- Conclusions

K-Means

- Unsupervised learning, clustering algorithm
- Clustering = organize objects into natural clusters

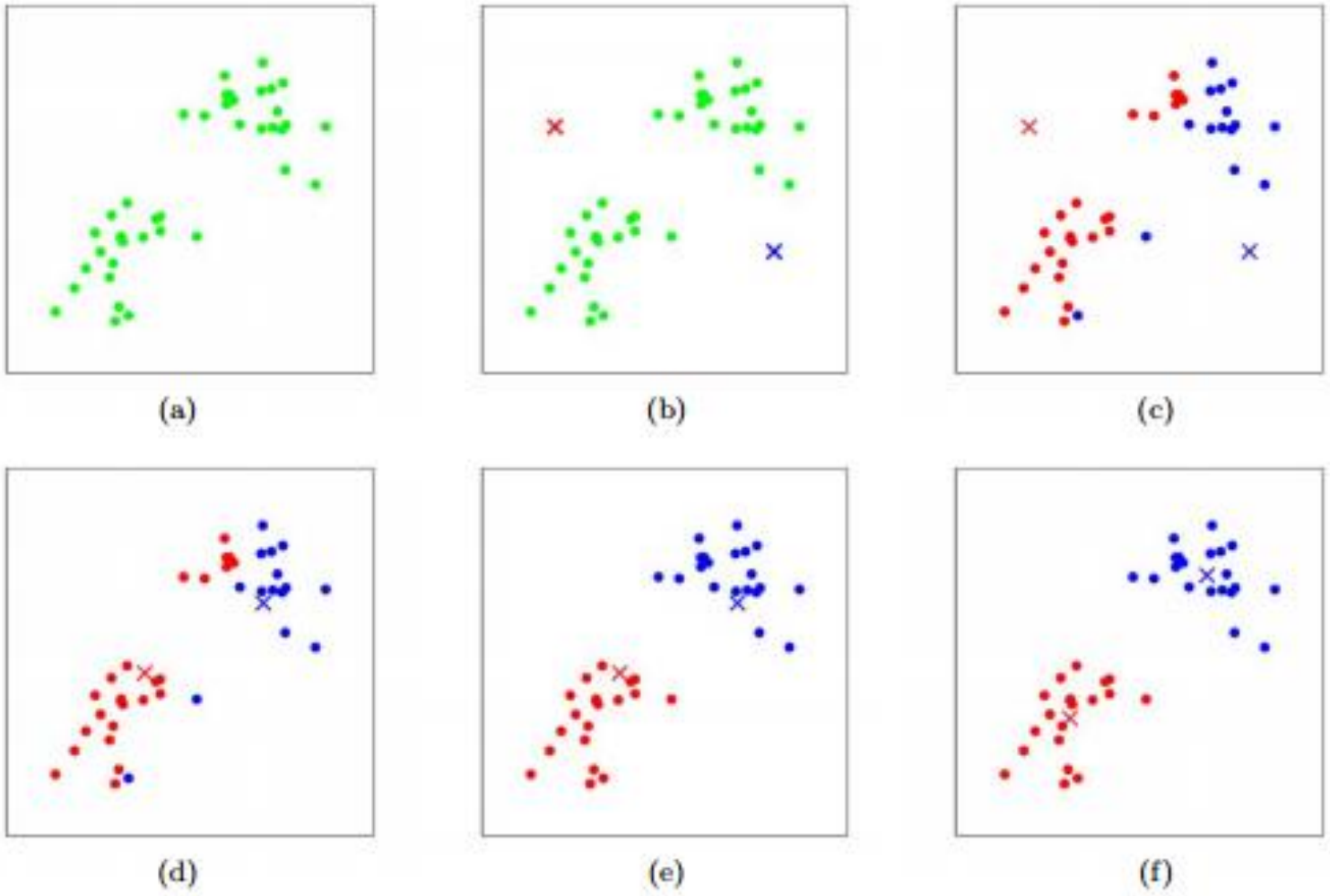
- K-Means is old but still popular because of its simplicity, efficiency and proven success

K-means : classic algorithm

Given N points, with D features, finds K clusters

1. Select an initial partition
2. Repeat 3 and 4 until a termination condition
3. Assign each point to the closest center
4. Update cluster centers based on the new assignment

K-means: visual example



Improvements

- Optimization of different aspects of the classic algorithm

1. Computation speedup
2. Seed selection optimization

1.1. Enhanced K-means

2006, Fahim, Salem, Torkey and Ramadan

- Memorize the distance between each point and the cluster centers
- At every iteration, if the distance to the new cluster mean is smaller => keeps the point in the assigned cluster
- Optimization for large values of K

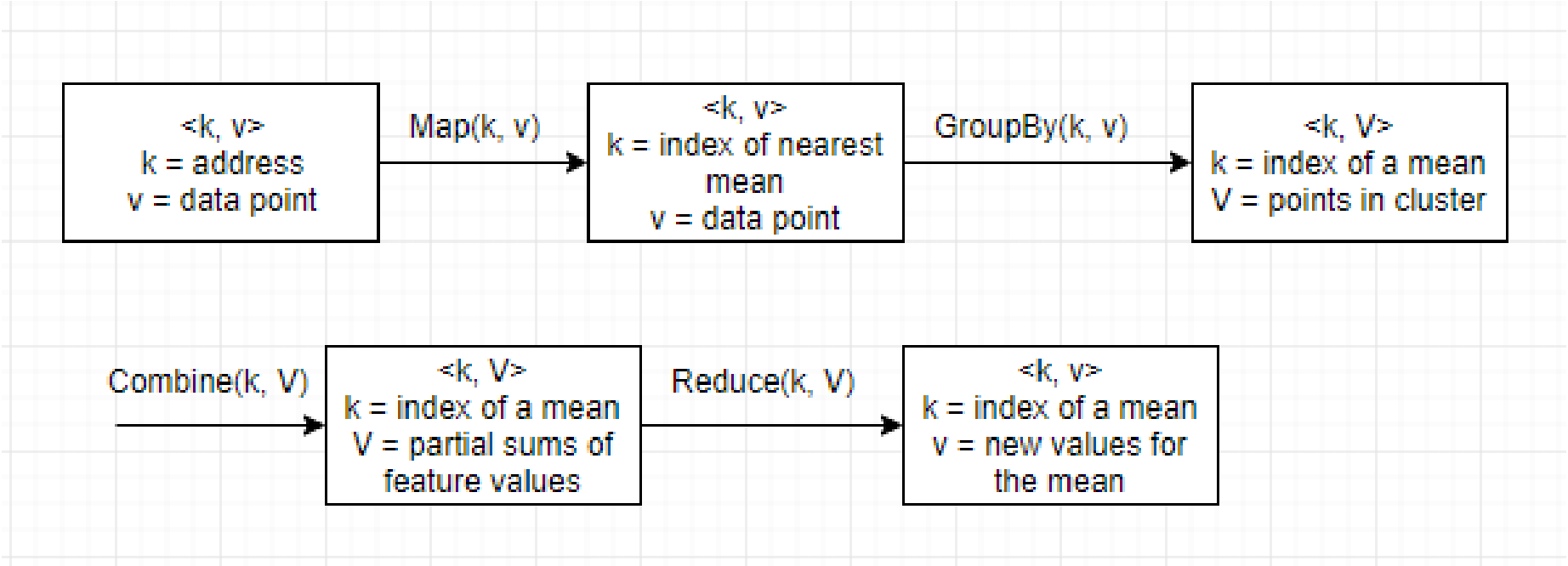
1.2. MapReduce K-Means

- Framework for processing large datsets over a cluster of computing nodes
- Popular in the Big Data community, with the Hadoop implementation
- Models problems as a combination of map and reduce functions

1.2. MapReduce K-Means

2009, W.Zhao, H. Ma, Q. He

- Data is stored in <key, value> pairs on a dfs
- Stores current centers in global variable



1.3. GPU K-Means

- GPU = special circuits designed for computer graphics and image processing
- Efficient at doing a simple operation over a large batch of data
- Can have thousands of threads

1.3. GPU K-means

published 2010, Li et al.

- Distance calculation: either dispatch each point to a thread or represent as matrixes and process in tiles data[n][d], centroid[d][k], result[n][k]
- Update means: divide and conquer; split data into M groups where M is the number of multiprocessors; reduce each group, split again until the dataset is < M

2.1. Sort and split heuristic

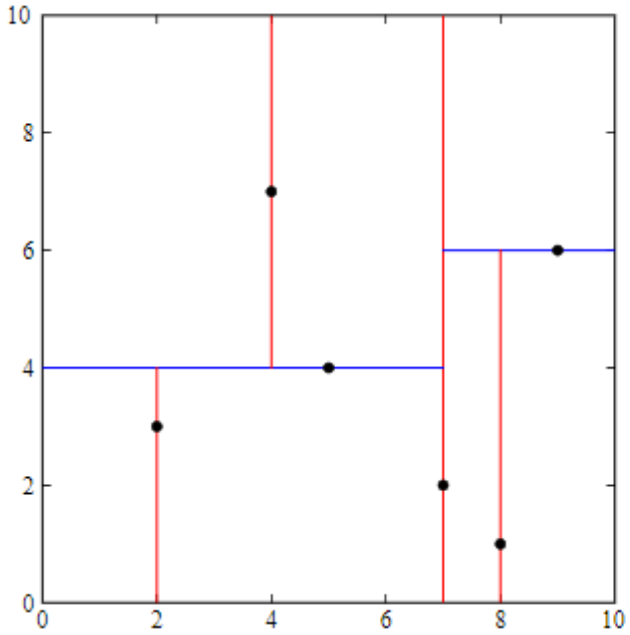
2010, Madhu and Pathakota

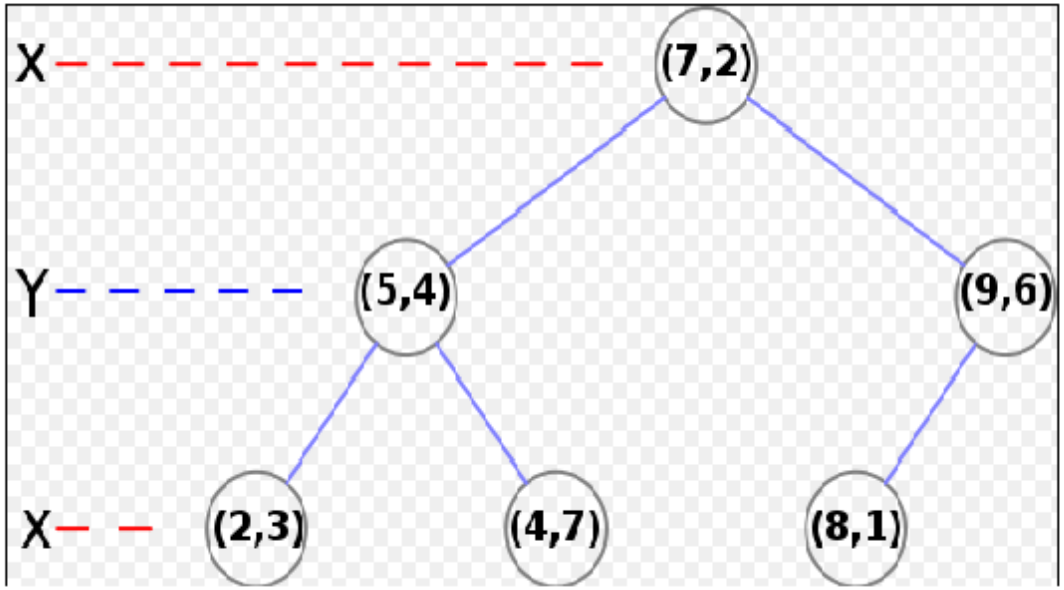
- Transform negative features by subtracting the minimum attribute value
- Sort all points by distance to origin
- Partition data into k equal sets
- Choose the middle point from each set as a seed

2.2. K-d trees

- Data structure for organizing points in k-dimensional space
- Built by recursively splitting the original dataset across a dimension
- A bounding box can be calculated in each subtree, to help with density estimations

2.2. K-d trees





2.2. K-d trees

2007, Redmond and Heneghan

- While building the tree: split over the largest dimension, leafs contain n/10k points
- Create density estimation for each leaf as : nr points / volume of bounding box
- Heuristic : choose means from buckets with large densities, separated by a reasonable distance

Variants

- Different approaches for the classic algorithm

1. Bisecting K-Means
2. GA K-Means

1. Bisecting K-Means

2000, Steinbach, Karypis and Kumar

- Can perform both flat and hierarchical clustering
- At each iteration splits the dataset into 2 by applying the basic algorithm
- Repeats until k clusters have been found
- More efficient when k is large, tends to produce balanced clusters

2. Genetic Algorithms

- Metaheuristic search method inspired by nature
- Models solution as chromosomes and scores them with a function called fitness
- Random mutations of solution can escape local optima

2.1. KGA-clustering

2002, Maulik and Bandiopadhyay

- Chromosome: array of cluster centers
- Fitness: distortion
- Crossover: single point
- Mutation: a random feature is updated with a fixed probability
- For each chromosome: decode, update means, apply crossover and mutation, add to next generation if fitness is good

2.2. IGKM

2006, Guo, Liu, Gao and Zhu

- Chromosomes also contain a value for k
- Fitness: (1/k) * (E1/Ek) * Dk
- Ek = distortion for k clusters
- Dk = maximum distance between 2 cluster centers
- Penalizes large k, increases for compact and separated clusters

Conclusions

- K-Means can handle just about any dataset
- Remains popular : around 30.000 results on Google Scholar each year