#### Report

**on**

#### K-means Clustering

# Rachita Bansal

**Loukik Kulkarni**

**Neharika Mandhan**

**Biplab Kumar Ghosh**

**Clustering**

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory [data mining](http://en.wikipedia.org/wiki/Data_mining), and a common technique for [statistical](http://en.wikipedia.org/wiki/Statistics) [data analysis](http://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [machine learning](http://en.wikipedia.org/wiki/Machine_learning), [pattern recognition](http://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](http://en.wikipedia.org/wiki/Image_analysis), [information retrieval](http://en.wikipedia.org/wiki/Information_retrieval), and [bioinformatics](http://en.wikipedia.org/wiki/Bioinformatics).

*Organizing data into clusters such that there is:*

→ High intra-cluster similarity

→ Low inter-cluster similarity

→ Informally, finding natural groupings among objects

*Purpose of clustering:*

→ Organizing data into clusters shows internal structure of the data Ex. Clustering genes

→ Sometimes the partitioning is the goal Ex. Market segmentation

→ Prepare for other AI techniques Ex. Summarize news (cluster and then find centroid)

→ Discovery in data Ex. Underlying rules, reoccurring patterns, topics, etc.

**Scenarios to implement Clustering**

* A telephone company needs to establish its network by putting its towers in a particular region it has acquired. The location of putting these towers can be found by clustering algorithm so that all its users receive maximum signal strength.
* Cisco wants to open its new office in California. The management wants to be cordial to its employees and want their office in a location so that its employees’ commutation is reduced to minimum.
* The Miami DEA wants to make its law enforcement more stringent and hence have decided to make their patrol vans stationed across the area so that the areas of high crime rates are in vicinity to the patrol vans.
* A Hospital Care chain wants to open a series of Emergency-Care wards, keeping in mind the factor of maximum accident prone areas in a region.

**K-means clustering**

Clustering is the process of partitioning a group of data points into a small number of clusters. For instance, the items in a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have *n* data points **x***i*,*i*=1...*n* that have to be partitioned in *k* clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions *μi*,*i*=1...*k* of the clusters that minimize the *distance* from the data points to the cluster. K-means clustering solves

argmin**c**∑*i*=1*k*∑**x**∈*cid*(**x**,*μi*) =argmin**c**∑*i*=1*k*∑**x**∈*ci*∥**x**−*μi*∥22

where **c***i* is the set of points that belong to cluster *i*. The K-means clustering uses the square of the Euclidean distance *d*(**x**,*μi*) =∥**x**−*μi*∥22. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

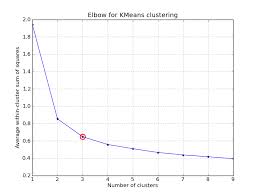
**K-mean algorithm**

The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters *k*. Then:

|  |  |
| --- | --- |
| 1. Initialize the center of the clusters | *μi*=[*some value*](http://www.onmyphd.com/?p=k-means.clustering#h3_init),*i*=1,...,*k* |
| 2. Attribute the closest cluster to each data point | **c***i*={*j*:*d*(**x***j*,*μi*)≤*d*(**x***j*,*μl*),*l*≠*i*,*j*=1,...,*n*} |
| 3. Set the position of each cluster to the mean of all data points belonging to that cluster | *μi*=1|*ci*|∑*j*∈*ci***x***j*,∀*i* |
| 4. Repeat steps 2-3 until convergence |  |
| Notation | |**c**|= number of elements in **c** |

### **Deciding the number of clusters**

### The number of clusters should match the data. An incorrect choice of the number of clusters will invalidate the whole process. An empirical way to find the best number of clusters is to try K-means clustering with different number of clusters and measure the resulting sum of squares.



One of the method to determine the number of clusters is the Elbow method. It looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified. Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an [F-test](http://en.wikipedia.org/wiki/F-test). A slight variation of this method plots the curvature of the within group variance. The method can be traced to speculation by [Robert L. Thorndike](http://en.wikipedia.org/wiki/Robert_L._Thorndike) in 1953.

### **Initializing the position of the clusters**

The initial position of the clusters is very important, there are few methods.

***Random***

* Assign each object to a random cluster.
* Computes the initial centroid of each cluster.

***Forgy***

* Chooses k objects at random and uses them as the initial centroids.

***Macqueen***

* Chooses k objects at random and uses them as the initial centroids.
* Assign each object to the cluster with the nearest centroid.
* After each assignment, recalculate the centroid.

***Kaufman***

* Select the most centrally located object as the first centroid.
* For every non-selected object *wi,*

Calculate *Cji = max (Dj-dji,0),*where *dji* is the distance between *wi* and *wj*, *Dj* is the distance between *wj* and its nearest centroid.

Calculate *∑jCji*

* Select *wi* which maximizes *∑jCji.*
* If there are *k* Centroids, stop. Otherwise go to step 2.

**Limitations**

* Applicable only when mean is defined, not on categorical data.
* Need to specify k, the number of clusters, in advance.
* Unable to handle noisy data and outliers.
* Not suitable to discover clusters with non-convex shapes.

**Applications**

* Optical Character Recognition

Optical character recognition has been an active area of research. In the first step, affine transformations are applied to the training samples in order to make the scheme robust against distortion. Scaling and Rotation are among those popular distortions which have been considered in this work. Inactive pixels are cut off using a simple algorithm in the next step. Then, principal component analysis (PCA) and k-means clustering are applied. The results from preprocessing showed a great potential in dimensionality reduction using transformations that can preserve useful information. Numerical results on the MNIST dataset reached 3% error rate which is lower than the other linear approaches. The proposed linear techniques are discussed in a way that make it easier to have a much clearer understanding of the method and why it works compared to the other classification methods.

* Biometrics

Biometric is a quantity which consists of individual physical characteristics that provide more authentication and security than the password, pin number, etc. The features of Fingerprint and (FKP) are extracted. The feature values of fingerprint using Discrete Wavelet Transform and the key points of FKP are clustered using K-Means clustering algorithm and their values are fused. The fused values of K-Means clustering algorithm is stored in a database which is compared with the query fingerprint and FKP K-Means centroid fused values to prove the recognition and authentication.

* Diagnostic Systems

Computer system supports diagnosis based on microscope images of the fine needle biopsy. The system assumes distinguishing malignant from benign cases. At first, thresholding reveals objects on background. Then image is clustered with k-means algorithm to distinguish nuclei from red blood cells and other objects. Correct segmentation is crucial to obtain good quality features measurements and consequently successful diagnosis. The system of malignancy classification was tested on a set of real case medical images with promising results.

* Commercial Application

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits and so on.

**Reference**

<http://en.wikipedia.org/wiki/Cluster_analysis>

<http://www.onmyphd.com/?p=k-means.clustering>

<http://en.wikipedia.org/wiki/Unsupervised_learning>

<http://en.wikipedia.org/wiki/K-means_clustering>

<http://ee.ucd.ie/~stephenr/documents/kmeans_elsart.pdf>

<http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set>

<http://www.bx.psu.edu/old/courses/bx-fall04/How_Many_Clusters.pdf>

**Dataset Details:**

<http://archive.ics.uci.edu/ml/datasets/Abalone>

***Title of Database***: Abalone data

***Sources***:

(a) Original owners of database:

Marine Resources Division

Marine Research Laboratories - Taroona

Department of Primary Industry and Fisheries, Tasmania

GPO Box 619F, Hobart, Tasmania 7001, Australia

(Contact: Warwick Nash +61 02 277277, wnash@dpi.tas.gov.au)

(b) Donor of database:

Sam Waugh (Sam.Waugh@cs.utas.edu.au)

Department of Computer Science, University of Tasmania

GPO Box 252C, Hobart, Tasmania 7001, Australia

***Relevant Information***:

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).

*Data comes from an original (non-machine-learning) study:*

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and

Wes B Ford (1994) "The Population Biology of Abalone (\_Haliotis\_

species) in Tasmania. I. Blacklip Abalone (\_H. rubra\_) from the North

Coast and Islands of Bass Strait", Sea Fisheries Division, Technical

Report No. 48 (ISSN 1034-3288)

***Number of Instances***: 4177

***Number of Attributes***: 8

***Attribute information***:

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Name Data Type Meas. Description

---- --------- ----- -----------

Sex nominal M, F, and I (infant)

Length continuous mm longest shell measurement

Diameter continuous mm perpendicular to length

Height continuous mm with meat in shell

Whole weight continuous grams whole abalone

Shucked weight continuous grams weight of meat

Viscera weight continuous grams gut weight (after bleeding)

Shell weight continuous grams after being dried

Rings integer +1.5 gives the age in years

***Missing Attribute Values***: None

***Analysis* *and Conclusion*:**

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task.

By applying K-means clustering to the dataset we intend to partition the whole data into a small number of clusters. In this case, based on the properties (Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight and Rings) of the abalone, we will make 4 clusters of data. Each cluster of abalone will have similar kind of properties.

In this particular dataset, in order to plot a graph with number of clusters in the x-axis and sum of squares in the y-axis, we have implemented the following code in R.

wssplot <- **function**(**data**, nc=15, seed=1234){

wss <- (**nrow**(**data**)-1)\***sum**(**apply**(**data**,2,**var**))

**for** (i **in** 2:nc){

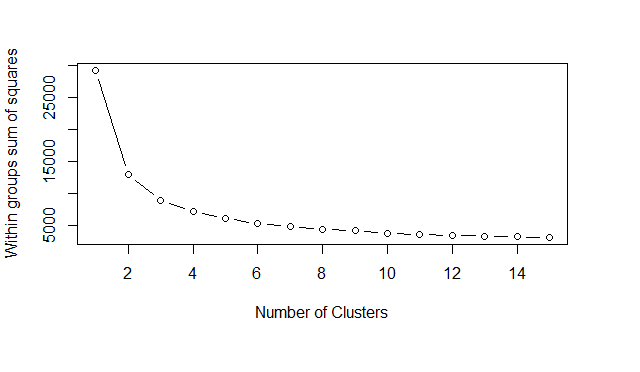
**set.seed**(seed)

wss[i] <- **sum**(**kmeans**(**data**, centers=i)$withinss)}

**plot**(1:nc, wss, type="b", xlab="Number of Clusters",

ylab="Within groups sum of squares")}

The data parameter is the numeric dataset to be analyzed, nc is the maximum number of clusters to be considered, and seed is a random number seed.



The bend in the graph suggests the number of clusters, which best suits the dataset, i.e. 4.

Implementing K-means Clustering on the dataset, which resulted in the formation of 4 different clusters. As per the KMC\_Output Cluster centers table1 (below), we can conclude, abalone in the cluster3 have all their properties on the higher side and that of cluster4 is on the lower end. At a glance, we can say, all the clusters of abalone are different from each other with a prominent difference in their properties.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Length** | **Diameter** | **Height** | **Whole weight** | **Shucked weight** | **Viscera weight** | **Shell weight** | **Rings** |
| **Cluster-1** | 0.595217 | 0.466863 | 0.159763 | 1.061964 | 0.461775 | 0.23225 | 0.303217 | 11.07249 |
| **Cluster-2** | 0.491432 | 0.38066 | 0.127529 | 0.583371 | 0.251673 | 0.127278 | 0.173194 | 9.461776 |
| **Cluster-3** | 0.677262 | 0.535111 | 0.192891 | 1.656155 | 0.716837 | 0.359135 | 0.468207 | 12.70238 |
| **Cluster-4** | 0.336846 | 0.253093 | 0.083594 | 0.201606 | 0.08812 | 0.04345 | 0.061683 | 6.635697 |

From the data summary table (below), we can determine maximum number of abalone have the properties of cluster1 and minimum number matches the properties of cluster3.

|  |  |  |
| --- | --- | --- |
| **Cluster** | **#Obs** | **Avg. Dist** |
| **Cluster-0** | 1476 | 1.218295 |
| **Cluster-1** | 1295 | 1.037334 |
| **Cluster-2** | 588 | 1.667236 |
| **Cluster-3** | 818 | 1.017403 |
| **Overall** | 4177 | 1.186048 |

Taking into account the major consideration of determining the age of the abalone, we can take one or two specimen from each of the clusters in order to predict the age of each abalone in the respective clusters.