IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 6 (Sep-Oct. 2012), PP 23-30
www.iosrjournals.org

A Comprehensive Overview of Clustering Algorithms in Pattern Recognition

Namratha M [1], Prajwala T R [2]
[1, 2](Dept. of information science, PESIT/visvesvaraya technological university, India)

Abstract: *Machine learning is a branch of artificial intelligence which recognizes complex patterns for making intelligent decisions based on input data values. In machine learning, pattern recognition assigns input value to given set of data labels.*
Based on the learning method used to generate the output we have the following classification, supervised and unsupervised learning. Unsupervised learning involves clustering and blind signal separation. Supervised learning is also known as classification.

This paper mainly focuses on clustering techniques such as K-means clustering, hierarchical clustering which in turn involves agglomerative and divisive clustering techniques.

This paper deals with introduction to machine learning, pattern recognition, clustering techniques. We present steps involved in each of these clustering techniques along with an example and the necessary formula used. The paper discusses the advantages and disadvantages of each of these techniques and in turn we make a comparison of K-means and hierarchical clustering techniques. Based on these comparisons we suggest the best suited clustering technique for the specified application.

Keywords: *Agglomerative, Clustering, Divisive, K-means, Machine learning, Pattern recognition*

I. Introudction

Machine learning is the field of research devoted to study of learning systems. Machine learning refers to changes in the systems that perform tasks associated with artificial intelligence like recognition, diagnosis, prediction and so on.

In machine learning[1], pattern recognition is assignment of label to given input value.

A pattern is an entity like fingerprint image, handwritten word or human face that could be given a name. Recognition is an act of associating a classification with a label.

Pattern recognition[2] is the science of making inferences based on data. It's objective is to assign an object or event to one of a number of categories based on features derived to emphasize commonalities.

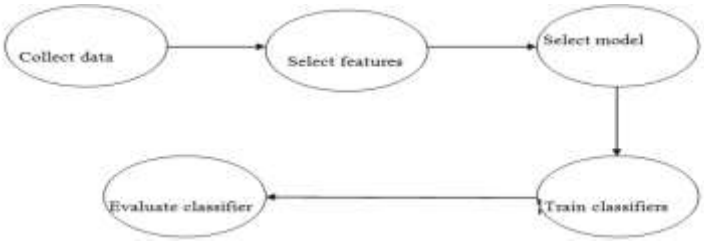


Figure 1: Design cycle for patter recognition

Pattern recognition involves three types of learning:

1. Unsupervised learning
2. Supervised learning
3. Semisupervised learning

In unsupervised learning[2] also known as cluster analysis, the basic task is to develop classification labels. It's task is to arrive at some grouping of data. The training set consists of labeled data.

Two types of unsupervised learning are:

a. Clustering
b. Blind signal separation

In supervised learning[2], classes are predetermined. The classes are seen as a finite set of data. A certain segment of data will be labeled with these classification. The task is to search for patterns and construct mathematical models. The training set consists of unlabeled data.

Two types of supervised learning are:

a. Classification
b. Ensemble learning

Semisupervised learning deals with methods for exploiting unlabeled data and labeled data automatically to improve learning performance without human intervention.

Four types of semisupervised learning are:

a. Deep learning
b. Low density separation
c. Graph based methods
d. Heuristic approach

Clustering is a form of unsupervised learning which involves the task of finding groups of objects which are similar to one another and different from the objects in another group. The goal is to minimize intracluster distances and maximize intercluster distances[3].

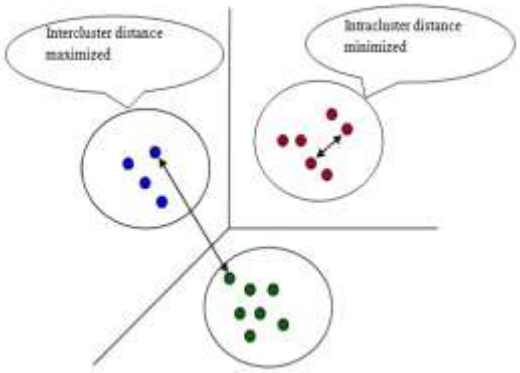


Figure 2: Graphical representation of clustering

II. K-Means Clustering

K-means is one of the simplest unsupervised learning algorithm that is used to generate specific number of disjoint and non-hierarchical clusters based on attributes[4]. It is a numerical, non-deterministic and iterative method to find the clusters. The purpose is to classify data.

Steps in K-means clustering:

Step 1: Consider K points to be clustered $x_1,\ldots, x_K$. These are represented in a space in which objects are being clustered. These points represent initial centroids.

Step 2: Each object is assigned to the group that has closest centroid[5].

$$m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k},\ k = 1,\ldots,K.$$

Step 3: The positions of K centroids are recalculated after all objects have been assigned. $C(i)$ denotes cluster number for the i^{th} observation[5]

$$C(i) = \arg\min_{1\le k\le K} \left\| x_i - m_k \right\|^2,\ i = 1,\ldots,N$$

Step 4: Reiterate steps 2 and 3 until no other distinguished centroid can be found. Hence, K clusters whose intracluster distance is minimized and intercluster distance is maximized[5].

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k}\left\| x_i - x_j \right\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k}\left\| x_i - m_k \right\|^2$$

where

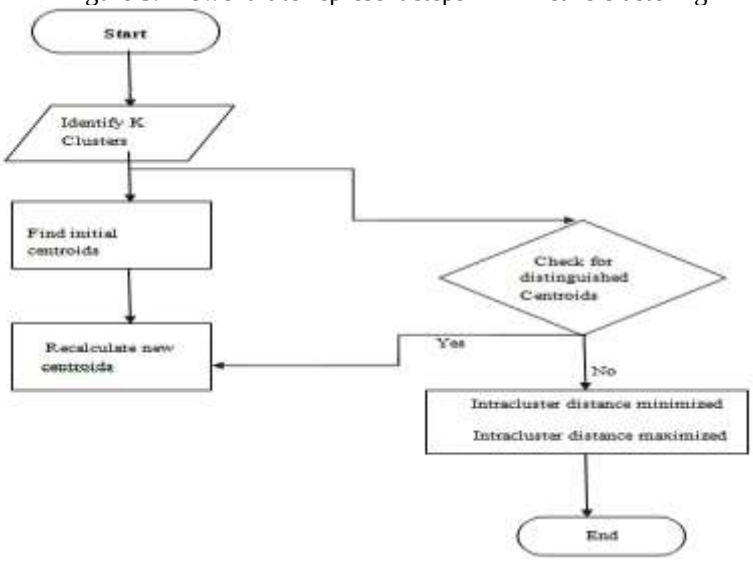
m_k is the mean vector of the k^{th} cluster

N_k is the number of observations in k^{th} cluster

The choice of initial cluster can greatly affect the final clusters in terms of intracluster distance, intercluster distance and cohesion.

The sum of squares of distance between object and corresponding cluster centroid is minimum in the final cluster.

Figure 3: Flowchart to represent steps in K-means clustering



Advantages:

1. K-means is computationally fast.
2. It is a simple and understandable unsupervised learning algorithm

Disadvantages:

1. Difficult to identify the initial clusters.
2. Prediction of value of K is difficult because the number of clusters is fixed at the beginning.
3. The final cluster patterns is dependent on the initial patterns.

Example: Problem:

To find the cluster of 5 points: (2,3),(4,6),(7,3),(1,2),(8,6).

Solution:

The initial clusters are (4,6) and (2,3) Iteration 1:

	(4,6)	(2,3)	Cluster
(2,3)	5	0	2
(1,2)	7	2	2
(4,6)	0	5	1
(8,6)	4	9	1
(7,3)	6	5	2

The cluster column is calculated by finding the shortest distance between the points[6]. The new values of centroid are (6,6) and (10/3,8/3).

Iteration 2:

Repeat the above steps to find the new values of centroids. Since the values converge, we do not proceed to next iteration. Hence the final clusters are :

Cluster 1: (4,6) and (8,6) Cluster 2:(2,3) , (1,2) and (7,3) **Applications**[7]:

- Used in segementation and retreival of grey level images[6].
- Applied for spatial and temporal datasets in the field of geostatics.
- Used to analyze the listed enterprises in financial organizations[4].
- Also used in the fields of astronomy and agriculture.

III. Hierarchical Clustering

It is an unsupervised learning technique that outputs a hierarchical structure which does not require to prespecify the nuimber of clusters. It is a deteministic algorithm[3].

There are two kinds of hierarchical clustering:
1. Agglomerative clustering
2. Divisive clusttering

Agglomerative clustering:

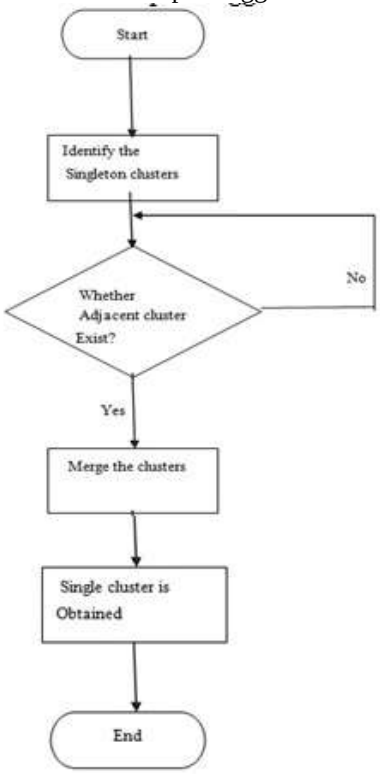
It is a bottom up approach with n singleton clusters initially where each cluster has subclusters which in turn have subclusters and so on[9].
Steps in agglomerative clustering:
Step 1: Each singleton group is assigned with unique data points.
Step 2: Merge the two adjacent groups iteratvely repeat this step. Calculate the Euclidian distance using the formula given below[8],

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Where a(x1,y1) and b(x2,y2) represent the coordinates of the clusters . Mean distance dmean(Di,Dj)=‖xi-xj‖
Where Di and Dj represent the clusters i and j respectively
Xi and xj are the means of clusters i and j respectively

Step 3: Repeat until a single cluster is obtained.

Figure 4: Flowchart for steps in agglomerative clustering



Advantages:
1. It ranks the objects for easier data display.
2. Small clusters are obtained which is easier to analyze and understand.
3. Number of clusters is not fixed at the beginning. Hence, user has the flexibility of choosing the clusters dynamically.

Disadvantages:
1. If objects are grouped incorrectly at the initial stages , they cannot be relocated at later stages.

2. The results vary based on the distance metrics used.

Example: Problem:
To find the cluster of 5 points: A(2,3),B(4,6),C(7,3),D(1,2),E(8,6).

Solution:
Iteration 1:
Calculate the Euclidian distance between two points. Euclidian distance between two points are:
A(2,3) and B(1,2)=sqrt(2)=1.41

A(2,3) and C(4,6)=sqrt(13)=3.6

A(2,3) and D(8,6)=sqrt(25)=5

A(2,3) and E(7,3)= sqrt(25)=5

The two adjacent clusters are A(2,3) and B(1,2). Merge these two clusters. The new centroid is F(1.5,2.5).
Iteration 2:
Repeat the above step and merge adjacent clusters as above.

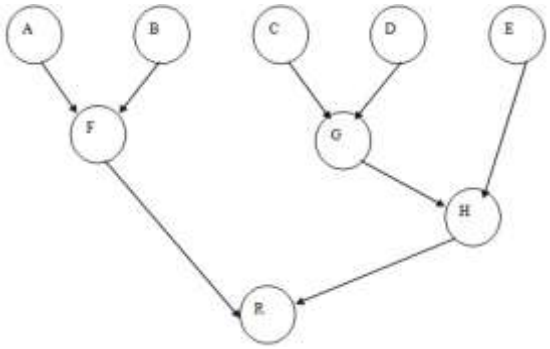
The two adjacent clusters are C(4,6) and D(8,6). Merge these two clusters. The new centroid is G(6,6).
Iteration 3:
Repeat the above step and merge adjacent clusters as above.

The two adjacent clusters are G(6,6) and E(7,3). Merge these two clusters. The new centroid is H(6.5,4.5).
Iteration 4:

Repeat the above step and merge adjacent clusters as above.

The two adjacent clusters are H(1.5,2.5) and F(6.5,4.5). Merge these two clusters. Finally we get the resultant single cluster R.

Figure 5: Diagrammatic representation of agglomerative clustering for the above example



Applications:
1. Used in search engine query logs for knowledge discovery.
2. Used in image classification systems to merge logically adjacent pixel values.
3. Used in automatic document classification.
4. Used in web document categorization.

Divisive clustering:

It is a top-down clustering method which works in a similar way to agglomerative clustering but in the opposite direction. This method starts with a single cluster containing all objects and then successively splits resulting clusters until only clusters of individual objects remain[10].

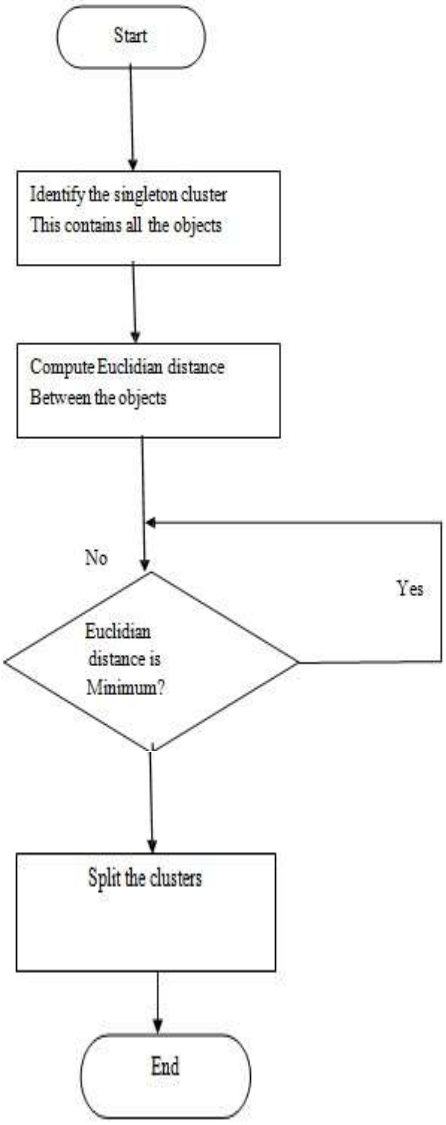
Steps in divisive clustering:
Step 1: Initially consider a singleton cluster.
Step 2: Iteratively divide the clusters into smaller clusters based on the Euclidian distance. Objects with least Euclidian distance are grouped into a single cluster[8].

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Where a(x1,y1) and b(x2,y2) represent the coordinates of the clusters .

Step 3: Repeat the process until desired number of clusters are obtained and Euclidian distance remains constant to obtain the final dendogram.

Figure 6: Flowchart for steps in divisive clustering



Advantages:

1. Focuses on the upper levels of dendogram.
2. We have access to all the data, hence the best possible solution is obtained.

Disadvantages:

1. Computational difficulties arise while splitting the clusters.
2. The results vary based on the distance metrics used.

Example: Problem:

To find the cluster of 5 points: A(2,3),B(4,6),C(7,3),D(1,2),E(8,6).

Solution:

Iteration 1: Calculate the Euclidian distance between two points. Euclidian distance between two points are:

	A(2,3)	B(4,6)	C(7,3)	D(1,2)	E(8,6)
A(2,3)		Sqrt(13)	Sqrt(25)	Sqrt(2)	Sqrt(45)
B(4,6)			Sqrt(18)	Sqrt(25)	Sqrt(16)
C(7,3)				Sqrt(37)	Sqrt(10)
D(1,2)					Sqrt(65)
E(8,6)					

Since sqrt(2) is the least Euclidian distance merge the point sA(2,3) and D(1,2)
The new centroid is F(1.5,2.5). Iteration 2:

Repeat the above step and merge adjacent clusters with least Euclidian distance as above. The two adjacent clusters are C(7,3) and E(8,6). Merge these two clusters.
The new centroid is G(7.5,4.5). Iteration 3:

Repeat the above step and merge adjacent clusters with least Euclidian distance as above.
The two adjacent clusters are B(4,6) and G(7.5,4.5). Merge these two clusters.

Figure 7: The resulting dendogram for the above example[11].

Applications:

1. Used in medical imaging for PET scans.
2. Used in world Wide Web in social networking analysis and slippy map optimization.
3. Used in market research for grouping shopping items.
4. Used in crime analysis to find hot spots where crime has occurred.
5. Also used in mathematical chemistry and petroleum geology.

Agglomerative versus divisive clustering:

Table 1:Comparison of hierarchical clustering techniques

Agglomerative	Divisive
1. Bottom-up approach	Top-down approach
2. Faster to compute	Slower to compute
3. More ration al to global structure of data.	Less blind to global structure of data
4. Best possible merge is obtained	Best possible split is obtained
5. Access to individual objects	Access to all data

IV. K-Means versus Hierarchical Clustering

Table 2:Comparison of clustering techniques

Hierarchical	K means
Sequential partitioning process	Iterative partitioning process
Results in nested cluster structure	Results in Flat mutually exclusive structure
Membership of an object or cluster in fixed	Membership of an object or cluster could be constantly changed.
Prior knowledge of the number of clusters is not needed.	Prior knowledge of the number of clusters is needed in advance.
Generic clustering technique irrespective of the data types.	Data are Summarized by representative entities.
Run time is slow.	Run time Faster than Hierarchical.
Hierarchical clustering requires only a similarity measure.	K-means clustering requires stronger assumptions such as number of clusters and the initial centers.

V. Conclusion

This paper discusses the clustering techniques along with an illustrative example. By comparing the advantages and disadvantages of each of these techniques we made a list of the applications where the techniques could be used. Whenever we require a sequential partitioning and time is not a constraint hierarchical clustering can be used. Contradictorily when prior knowledge of clusters is available and mutually exclusive structure is used as training data we use K-means clustering. Each of the techniques described in this paper has it's own advantages and disadvantages. To overcome these disadvantages optimization techniques can be used for better performance.

References

[1] Tom Mitchell: "Machine Learning", McGraw Hill, 1997.
[2] http://www.springer.com/computer/image+processing/book/978-0-387-31073-2
[3] Data Clustering: A Review A.K. JAIN Michigan State University M.N. MURTY Indian Institute of Science AND P.J. FLYNN The Ohio State University
[4] http://books.ithunder.org/NLP/%E6%90%9C%E7%B4%A2%E8%B5%84%E6%96%99/%E6%96%87%E6%9C%AC%E8%81%9A%E7%B1%BB/k-means/kmeans11.pdf
[5] http://gecco.org.chemie.uni-frankfurt.de/hkmeans/H-k-means.pdf
[6] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html#macqueen
[7] http://delivery.acm.org/10.1145/2350000/2345414/p106-mishra.pdf?ip=119.82.126.162&acc=ACTIVE%20SERVICE&CFID=109187129&CFTOKEN=83079467&__acm__=1346224103_195980e7451e402acb712a927456104d
[8] http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/bishop-regression.pdf
[9] http://delivery.acm.org/10.1145/2010000/2003657/p34-spiegel.pdf?ip=119.82.126.162&acc=ACTIVE%20SERVICE&CFID=109187129&CFTOKEN=83079467&__acm__=1346223866_ec4f1d23636d3f275175a2f7bc11c432
[10] http://www.frontiersinai.com/ecai/ecai2004/ecai04/pdf/p0435.pdf
[11] http://delivery.acm.org/10.1145/950000/944973/3-1265-dhillon.pdf?ip=119.82.126.162&acc=PUBLIC&CFID=109187129&CFTOKEN=83079467&__acm__=1346223975_b9279bd748e1660bad277d28c67683cc