International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657
Volume 3, Issue 1, Feb 2015

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 3, Issue 1, Feb 2015

Comprehensive Overview of Clustering Algorithms in Pattern Recognition

ch.kalyani

Abstract

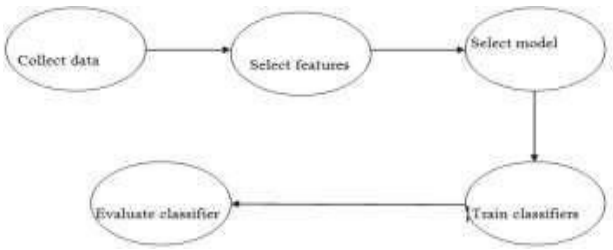
Machine learning is a subfield of AI that uses pattern recognition to evaluate data and derive meaningful conclusions. Pattern recognition is used in machine learning to provide an output value to a series of data labels. It is possible to categorize learning strategies according to the means by which final results are generated: supervised and unsupervised. Clustering and blind signal separation are at the heart of unsupervised learning. Classification is another name for supervised learning.K-means clustering, hierarchical clustering, and the agglomerative and divisive clustering methods they entail are the primary emphasis of this work.This article provides a primer on the fundamentals of machine learning, pattern recognition, and clustering methods. We detail the methodology behind each of these clustering methods and provide an example and the relevant formula. The study compares and contrasts K-means and hierarchical clustering methods, discussing the benefits and drawbacks of each. Using these contrasts, we advise which clustering method is ideal for the given task.

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

Introduction

The study of learning systems is the academic discipline known as "machine learning." The term "machine learning" is used to describe the adaptation of systems to improve their performance in AI-related activities such as recognition, diagnosis, prediction, and so on.attern recognition in machine learning[1] is the process of applying a label to an input value.A pattern is something that can be named, much as a fingerprint is something that can be used to identify a person. To recognize something is to give it a name hat fits that category.Data-driven inference is the basis of pattern recognition[2]. Its purpose is to classify things and events into distinct buckets according to their shared characteristics.

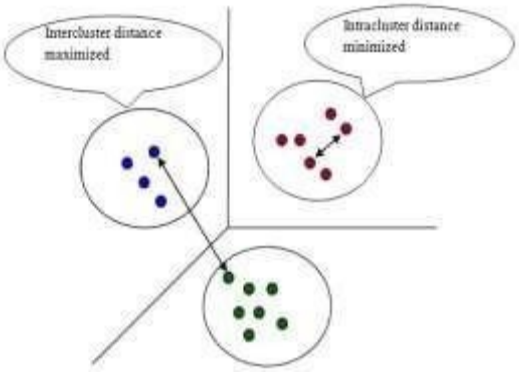
Pattern recognition involves three types of learning: UnsupervisedlearningSupervised rningSemisupervised learningIn unsupervised learning[2] also knownas cluster analysis, the basic task is to develop classification labels. It's task is to arrive at some grouping of data. The training set consists of labeled data



Department of
information technology

ISSN 2347–3657
Volume 3, Issue 1, Feb 2015

Two types of unsupervised learning are:
ClusteringBlind signal separationIn 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:Classification



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:
Deep learning
Low density separation
Graph based methods
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
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 $x1,\ldots, xK$. 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].

$$\sum x_i$$
$$m_k \quad =i:C(i)=k$$
$$N,\ k=1,\ldots,K.$$
$$k$$

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]

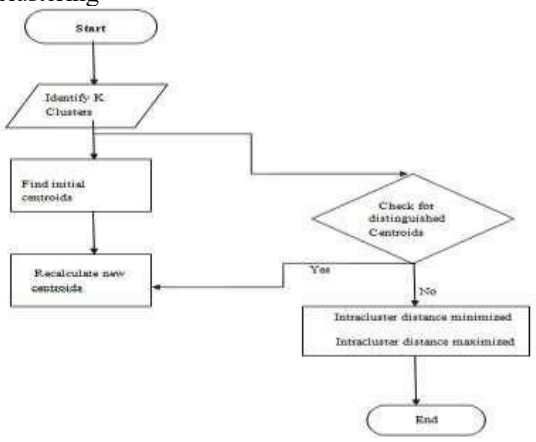
$$C(i)=\arg\min$$
$$x_i-m_k^2,\ i=1,\ldots,N$$
$$1\le k\le K$$

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].

$$(C)=\frac{1}{\sum\sum x_i-x_j 2}{}^{K}=\sum^{K}$$
$$N_k\ \sum x_i-\overset{2}{m_k}\|$$

where
$=1\ C(i)=k\ C(j)=kk=1C(i)=k$
is the mean vector of the k^{th} clusterNk is the number of observations in k^{th} clusterThe 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:

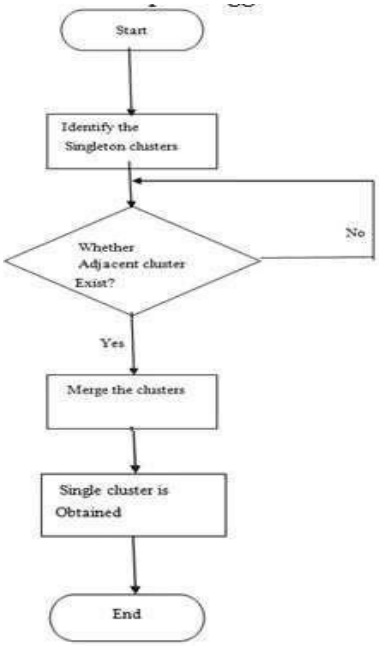
ISSN 2347–3657
Volume 3, Issue 1, Feb 2015

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

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

Disadvantages:
Difficult to identify the initial clusters. Prediction of value of K is difficult because the number of clusters is fixed at the beginning.
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:
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).

	(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

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.
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:
Agglomerative clustering
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],

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

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

ISSN 2347–3657
Volume 3, Issue 1, Feb 2015

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.

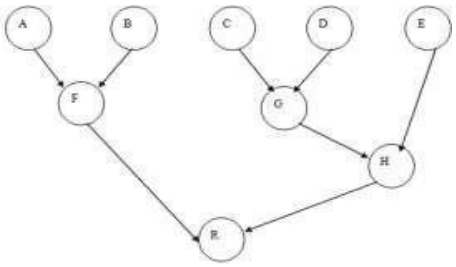
The two adjacent clusters are C(4,6) and D(8,6). Merge these two clusters. The new centroid is G(6,6).
Iteration 3:
Repeat the above step and merge adjacent clusters as above.

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:
Used in search engine query logs for knowledge



discovery.

Used in image classification systems to merge logically adjacent pixel values.
Used in automatic document classification.
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].

	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)					

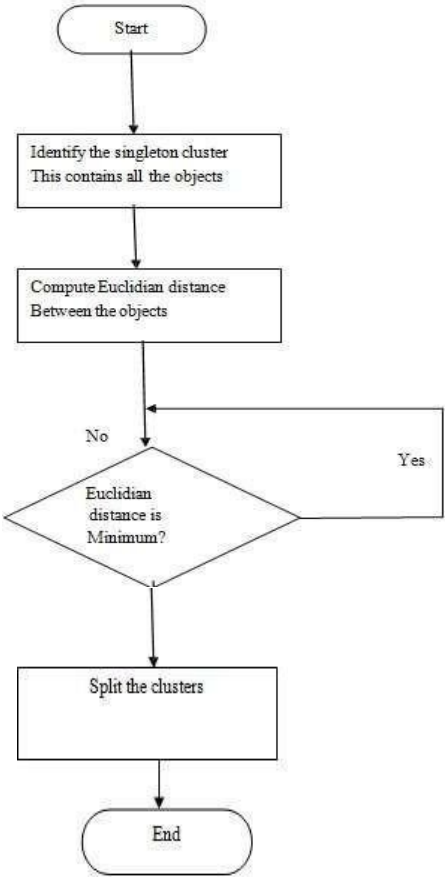
Steps in divisive clustering:
Step 1: Initially consider a singleton cluster.
Step 2: Iteratively divide the clusters into smaller clusters based on the Euclidian distance. Objects with least Euclidian distance are grouped into a single cluster[8].

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

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

International Journal of
Information Technology & Computer Engineering
ISSN 2347–3657
Volume 3, Issue 1, Feb 2015



AdvantagesFocuses on the upper levels dendogram.We have access to all the data, hence the best possible solution is obtained.

Disadvantages:
Computational difficulties arise while splitting the clusters.
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: 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:
Used in medical imaging for PET scans.
Used in world Wide Web in social networking analysis and slippy map optimization.
Used in market research for grouping shopping items.
Used in crime analysis to find hot spots where crime has occurred.
Also used in mathematical chemistry and petroleum geology.

Agglomerative versus divisive clustering:

K-Means versus Hierarchical Clustering

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-meansclustering requires stronger assumptions such as number of clusters and the initial centers.

Table 1:Comparison of clustering techniques

Conclusion
This article provides a discussion of clustering methods and an example to demonstrate these methods. We compiled a list of potential uses for these methods by weighing the benefits and drawbacks of each. When time is not a concern and a sequential segmentation is necessary, hierarchical clustering is an appropriate method to apply. K-means clustering is employed when we have previous knowledge of clusters and the data has a mutually exclusive structure. There are benefits and drawbacks to using

ISSN 2347–3657
Volume 3, Issue 1, Feb 2015

any of the methods discussed above. Better performance may be achieved by using optimization strategies to counteract these drawbacks.

References

Tom Mitchell: "Machine Learning", McGraw Hill, 1997.
http://www.springer.com/computer/image+processing/book/978-0-387-31073-2
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
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
http://gecco.org.chemie.uni-frankfurt.de/hkmeans/H-k-means.pdf
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html#macqueen
http://delivery.acm.org/10.1145/2350000/2345414/p1066-mishra.pdf?ip=119.82.126.162&acc=ACTIVE%20SERVICE&CFID=109187129&CFTOKEN=83079467&_a cm =1346224103_195980e7451e402acb712a927456104d
http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/bishop-regression.pdf
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_ec4f1d23636d3f275175a2f7bc11e432
http://www.frontiersinai.com/ecai/ecai2004/ecai04/pdf/p0435.pdf
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