**Fuzzy Clusters**

Given a set of data points, traditional clustering techniques partition the data into several groups such that the degree of association is strong within one group and weak between data in different groups. Similarity is high among the points in the intra clusters and low among points in the inter clusters. Classical clustering techniques result in partitions where each data point can belong to only one cluster. Fuzzy clustering by contrast allows data points to belong to more than one group. The resulting partition is therefore a fuzzy partition. Each cluster is associated with a membership function that expresses the degree to which individual data points belong to the cluster

**Fuzzy Clustering**

Fuzzy clustering is also known as soft clustering,because it allows an object to belong to more than one cluster. Consider the following scenario:

An online electronics store called MMcart, records customer’s browsing behaviour in a log. Here the data mining task is to use the log data of customers to classify them based on their search intent. In the entire span of the time he spent in that online store, he/she must have browsed information about a particular product, or might have searched for customer service information. It is difficult to know the customer’s search intent in advance. As this problem sounds like an unsupervised learning, a clustering analysis helps. Here a cluster is one that contains similar user browsing activities.

Let each session be customer’s time spent in browsing. Sometimes, it may so happen that not every session, belongs to only one cluster. For example, suppose user sessions involving the purchase of mobile phones form one cluster, and user sessions that compare the price of laptop computers form another cluster. What if a user, in one session, makes an order for a phone, and simultaneously compares several laptop computers? Such a session belongs to both clusters .These type of clusters are called fuzzy clusters. So Fuzzy clusters provide the flexibility of allowing an object to participate in multiple clusters*.*

**Fuzzy Set**

Given a set of objects, X= {x1,x2,...xn} a fuzzy set *S* is a subset of *X* that allows each object in *X* to have a membership degree between 0 and 1. In general, a fuzzy set, *S*, can be defined as a function, Fs: X -> [0,1]

The particular brand of a mobile phone is more popular if more units are sold. The degree of popularity of a mobile phone, O, is measured by the number of sales of it. To compute the degree of popularity of a phone, the following formula is used.

Popularity (O) 1 if 1000 or more units of O are sold

i /1000 if i ( i < 1000) units of O are sold.

Function Popularity () defines a fuzzy set of popular mobile phones. For example, the sales of mobile phones at MMCartare as shown in the table. The fuzzy set of popular mobile phones is { A (.07), B(1), C(.89), D(.36) } where the degrees of membership are written in parentheses.

|  |  |
| --- | --- |
| ***Phones*** | ***Sales (units)*** |
| ***A*** | **70** |
| ***B*** | **1320** |
| ***C*** | **890** |
| ***D*** | **360** |

Suppose the MMCart online store has six reviews. The keywords contained in these reviews are listed in Table. We can group the reviews into two fuzzy clusters, *C*1 and *C*2. *C*1 is for “Mobile Phone” and “selfie stick ," and *C*2 is for “computer.”

The partition matrix

|  |  |
| --- | --- |
| *Review\_ID* | Keywords |
| *R1* | Mobilephone, selfie stick |
| *R2* | Mobile phone |
| *R3* | Selfie stick |
| *R4* | Mobile phone Selfie stick,  computer |
| *R5* | computer, CPU |

*1, 0*

*1, 0*

*M = 1, 0*

*2/3, 1/3*

*0, 1*

Here, the keywords “mobile phone” and “selfie stick” denote the features of cluster *C*1, and “computer” the feature of cluster *C*2. For review, *Ri*, and cluster, *Cj* (1<= i<=5, 1<=j<=2)

In this fuzzy clustering, review *R*4 belongs to clusters *C*1 and *C*2 with membership degrees 2/3 and 1/3, respectively.

The Sum of Squared Error (SSE) can be used to measure how well a fuzzy clustering fits a data set. Fuzzy clustering is also called soft clusteringbecause it allows an object to belong to more than one cluster.

The *k*-means clustering can be considered as a special case of fuzzy clustering

**Implementation in R Programming Language**

Let us take iris data set which has 3 classes.

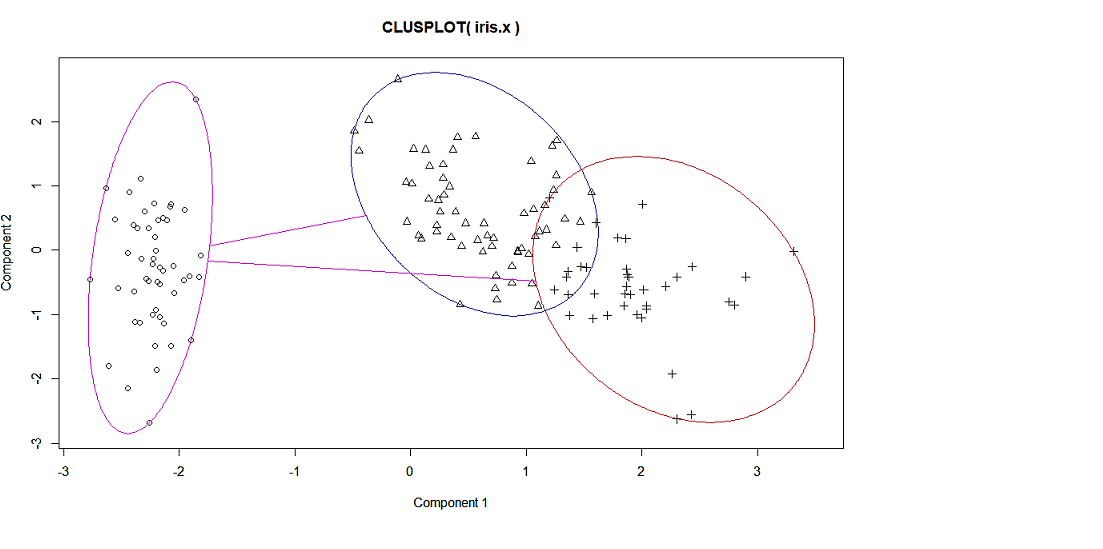
library(cluster)

data(iris)

iris.x <- iris[, 1:4]

cl3 <- pam(iris.x, 3)$clustering

clusplot(iris.x, cl3, color = TRUE)



The three clusters formed are shown in the above figure.

|  |
| --- |
| > fanny\_iris<-fanny(iris.x,3)  > fanny\_iris  Fuzzy Clustering object of class 'fanny' :  m.ship.expon. 2  objective 45.07716  tolerance 1e-15  iterations 28  converged 1  maxit 500  n 150  Membership coefficients (in %, rounded):  [,1] [,2] [,3]  [1,] 91 4 5  [2,] 86 6 8  [3,] 87 5 8  [4,] 84 7 9  [5,] 90 4 6  [6,] 77 10 13  [7,] 86 6 8  [8,] 92 3 4  [9,] 77 10 13  [10,] 88 5 7  [11,] 83 7 10  [12,] 89 4 6  [13,] 85 6 9  [14,] 75 11 14  [15,] 70 13 17  [16,] 64 15 20  [17,] 79 9 12  [18,] 91 4 5  [19,] 72 12 16  [20,] 86 6 8  [21,] 83 7 10  [22,] 87 5 7  [23,] 80 9 11  [24,] 85 6 9  [25,] 82 7 11  [26,] 85 6 9  [27,] 90 4 6  [28,] 90 4 6  [29,] 90 4 6  [30,] 87 5 8  [31,] 87 5 8  [32,] 84 7 9  [33,] 77 10 13  [34,] 72 12 16  [35,] 88 5 7  [36,] 87 5 7  [37,] 82 8 10  [38,] 89 5 6  [39,] 78 9 13  [40,] 92 3 5  [41,] 90 4 6  [42,] 66 14 20  [43,] 80 8 11  [44,] 85 6 9  [45,] 78 9 13  [46,] 85 6 9  [47,] 85 6 9  [48,] 86 6 8  [49,] 86 6 8  [50,] 92 4 5  [51,] 10 48 42  [52,] 8 35 57  [53,] 8 54 38  [54,] 13 22 65  [55,] 7 37 56  [56,] 7 20 73  [57,] 7 41 52  [58,] 30 24 47  [59,] 8 37 55  [60,] 15 23 62  [61,] 26 25 49  [62,] 7 20 72  [63,] 13 25 62  [64,] 6 28 66  [65,] 17 22 61  [66,] 9 37 53  [67,] 8 23 69  [68,] 10 20 70  [69,] 9 31 60  [70,] 13 21 66  [71,] 7 39 53  [72,] 9 21 70  [73,] 7 41 52  [74,] 7 28 66  [75,] 8 27 65  [76,] 9 35 57  [77,] 8 47 45  [78,] 6 60 34  [79,] 6 23 72  [80,] 20 23 58  [81,] 15 22 63  [82,] 18 22 60  [83,] 11 19 70  [84,] 7 44 49  [85,] 10 25 65  [86,] 9 33 59  [87,] 8 46 46  [88,] 9 29 62  [89,] 10 20 70  [90,] 12 21 68  [91,] 9 21 69  [92,] 6 27 67  [93,] 10 19 71  [94,] 29 24 48  [95,] 8 18 73  [96,] 9 19 72  [97,] 8 18 74  [98,] 7 23 70  [99,] 33 23 44  [100,] 8 18 74  [101,] 8 65 27  [102,] 7 46 47  [103,] 7 71 22  [104,] 5 69 26  [105,] 6 74 21  [106,] 12 59 29  [107,] 15 30 55  [108,] 10 63 27  [109,] 7 66 27  [110,] 10 63 27  [111,] 5 66 28  [112,] 5 66 29  [113,] 5 76 19  [114,] 9 42 50  [115,] 9 51 40  [116,] 6 69 25  [117,] 5 73 22  [118,] 14 56 31  [119,] 13 56 31  [120,] 9 39 53  [121,] 6 73 21  [122,] 9 39 52  [123,] 12 58 30  [124,] 6 48 46  [125,] 6 74 20  [126,] 8 67 25  [127,] 6 42 52  [128,] 6 45 48  [129,] 5 72 22  [130,] 8 66 26  [131,] 9 64 27  [132,] 14 56 31  [133,] 5 72 23  [134,] 6 50 44  [135,] 8 53 39  [136,] 11 61 28  [137,] 7 67 26  [138,] 5 71 24  [139,] 6 39 54  [140,] 5 74 21  [141,] 6 74 21  [142,] 7 67 26  [143,] 7 46 47  [144,] 7 72 21  [145,] 7 70 23  [146,] 6 70 24  [147,] 7 51 42  [148,] 5 71 25  [149,] 7 64 28  [150,] 7 48 45  Fuzzyness coefficients:  dunn\_coeff normalized  0.5679133 0.3518699  Closest hard clustering:  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  [46] 1 1 1 1 1 2 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 3 3 3 3 3 3 3 3 3 3 3 3  [91] 3 3 3 3 3 3 3 3 3 3 2 3 2 2 2 2 3 2 2 2 2 2 2 3 2 2 2 2 2 3 2 3 2 2 2 2 3 3 2 2 2 2 2 2 2  [136] 2 2 2 3 2 2 2 3 2 2 2 2 2 2 2  Available components:  [1] "membership" "coeff" "memb.exp" "clustering" "k.crisp" "objective"  [7] "convergence" "diss" "call" "silinfo" "data" |
| > library(factoextra)  > fviz\_cluster(fanny\_iris, ellipse.type = "norm", repel = TRUE,  + palette = "jco", ggtheme = theme\_minimal(),  + legend = "right") |
| C:\Users\malat\Desktop\analyticjoint\Rplot.jpeg  From the figure above we can find some data points that exhibit fuzziness.  [51,] 10 48 42  [77,] 8 47 45  [87,] 8 46 46  [127,] 6 42 52  [128,] 6 45 48  [150,] 7 48 45 |

> fviz\_silhouette(fanny\_iris, palette = "jco",

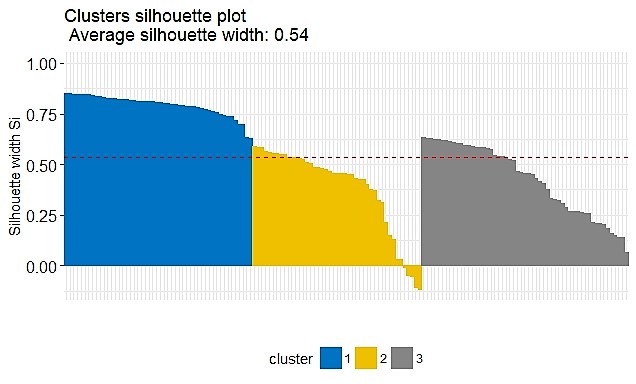
+ ggtheme = theme\_minimal())

cluster size ave.sil.width

1 1 50 0.79

2 2 45 0.38

3 3 55 0.43

****

The clusters and their size are given in the figure above.

For some applications like medicine and study of gene patterns in Bio-informatics fuzzy clustering is more appropriate than hard clustering.