Data Mining

Clustering III - Cluster Evaluation (Part A)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Cluster Evaluation

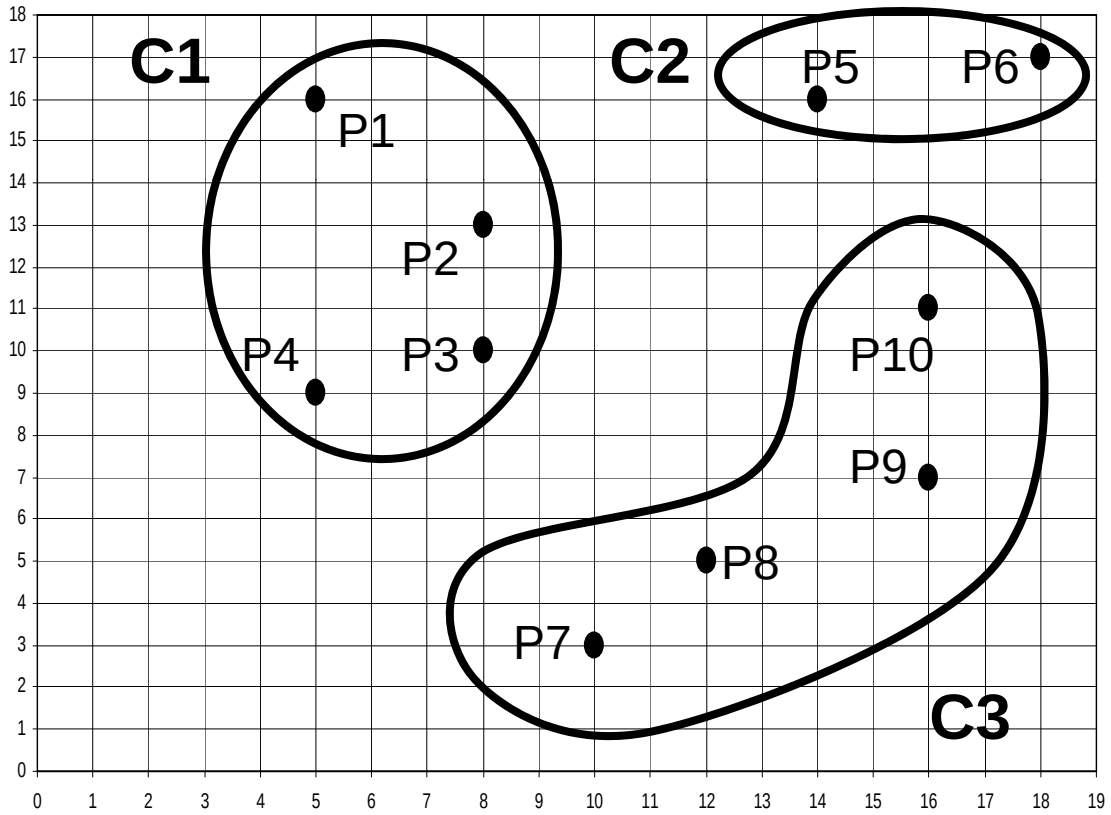
Cluster evaluation (or cluster validation)

- Need criteria for the quality of a clustering
- Need performance measures to evaluate how well a clustering algorithm works and to compare different clustering algorithms (which algorithm is better than another algorithm)

Silhouette coefficient - a good evaluation measure for clustering points in Euclidean space

Purity – a simple evaluation measure

The Silhouette Coefficient



The Silhouette Coefficient of P1

There are four points, P1(5, 16), P2(8, 13), P3(8, 10), P4(5, 9), in cluster C1.
Average distance from P1 to all other points in C1
= (dist(P1, P2) + dist(P1, P3) + dist(P1, P4))/3
= (4.24 + 6.71 + 7)/3
= 5.98
= $a1$

There are two points, P5(14, 16), P6(18, 17), in cluster C2.
Average distance from P1 to all points in C2
= (dist(P1, P5) + dist(P1, P6))/2
= (9 + 13.01)/2
= 11
= $m1$

There are four points, P7(10, 3), P8(12, 5), P9(16, 7), P10(16, 11), in cluster C3.
Average distance from P1 to all points in C3
= (dist(P1, P7) + dist(P1, P8) + dist(P1, P9) + dist(P1, P10))/4
= (13.93 + 13.04 + 14.21 + 12.08)/4
= 13.32
= $n1$

$b1$ = min($m1$, $n1$) = 11
$s1$ = ($b1$ – $a1$)/max($a1$, $b1$) = 5.02/11 = 0.46 where $s1$ is the silhouette coefficient of P1.

Average Silhouette Coefficient

There are four points, P1(5, 16), P2(8, 13), P3(8, 10), P4(5, 9), in cluster C1.
Average distance from P1 to all other points in C1
= 5.98
= $a1$

There are two points, P5(14, 16), P6(18, 17), in cluster C2.
Average distance from P1 to all points in C2
= 11
= $m1$

There are four points, P7(10, 3), P8(12, 5), P9(16, 7), P10(16, 11), in cluster C3.
Average distance from P1 to all points in C3
= 13.32
= $n1$

$b1$ = min($m1$, $n1$) = 11
$s1$ = ($b1$ – $a1$)/max($a1$, $b1$) = 5.02/11 = 0.46 where $s1$ is the silhouette coefficient of P1.
In general, the value of the silhouette coefficient varies between -1 ($b1$ = 0) and 1 ($a1$ = 0). We want the silhouette coefficient to be close to 1.

To measure the goodness of a clustering, we calculate the average silhouette coefficient of all points. If the average silhouette coefficient is 1, the clustering is perfect.

End of
Cluster Evaluation Module (Part A)