Scut: A Balanced Clustering Algorithm

Mikko Malinen

Speech and Image Processing Unit
School of Computing
University of Eastern Finland

2013

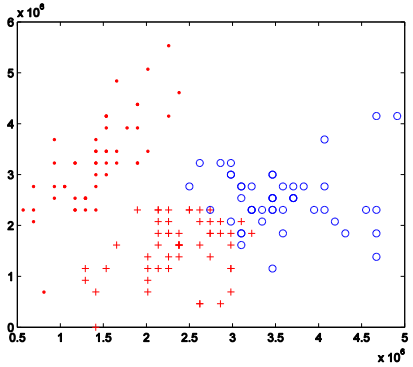
UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

A more balanced clustering result is desirable in some applications (f.eg. in photo query systems, workload balancing)



$$TSE_i = \sum_{\substack{j,\text{ so that} \\ X_i \text{ belongs} \\ \text{to cluster } i}} | X_j - C_i |^2$$

We present a clustering method, which weights TSE of a single cluster by number of points in cluster and uses those weighted TSE's as a cost.

$$\text{Cost} = n_1 \cdot TSE_1 + n_2 \cdot TSE_2 + \ldots + n_k \cdot TSE_k,$$

where n_i is number of points in ith cluster.

UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

The cost function

$$\text{Cost} = n_1 \cdot TSE_1 + n_2 \cdot TSE_2 + \ldots + n_k \cdot TSE_k,$$

where n_i is number of points in ith cluster.

may also be written as

$$\text{Cost} = {n_1}^2 \cdot MSE_1 + {n_2}^2 \cdot MSE_2 + \ldots + {n_k}^2 \cdot MSE_k,$$

where n_i is number of points in ith cluster.

$$(MSE_i = \sum_{\substack{j, \text{so that} \\ X_j \text{ belongs} \\ \text{to cluster } i}} \frac{|X_j - C_i|^2}{n_i})$$

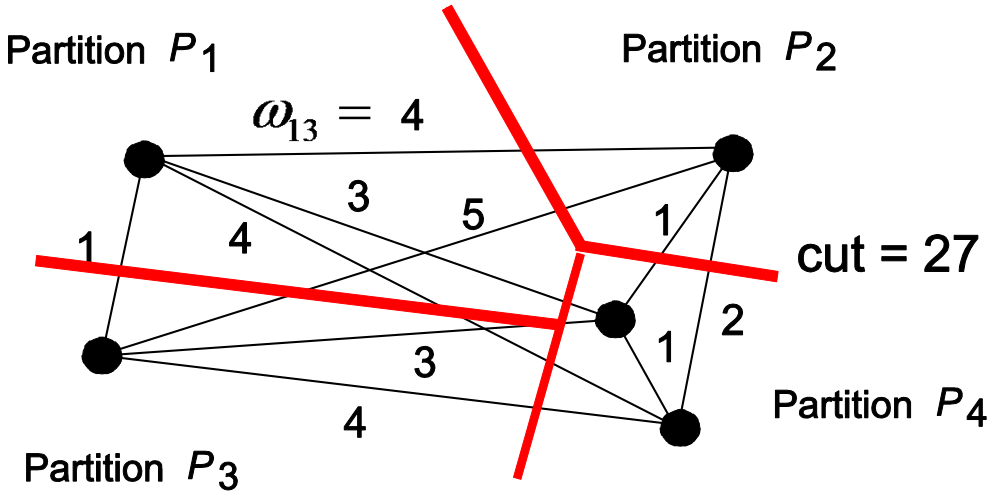
UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

MAX k-CUT



MAX 4-CUT

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Scut algorithm

Input: dataset X, number of clusters k

Output: partitioning of points P

Set the vertices of the graph the data points.

Set the weights of the edges of the graph the squared distances.

Compute MAX k-CUT

Output partitioning of points P

UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

UNIVERSITY OF
EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Balanced k-means clustering

Mikko Malinen

Speech and Image Processing Unit
School of Computing
University of Eastern Finland

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Balanced clustering

Want to fix the number of points in clusters?

We have a solution for that.

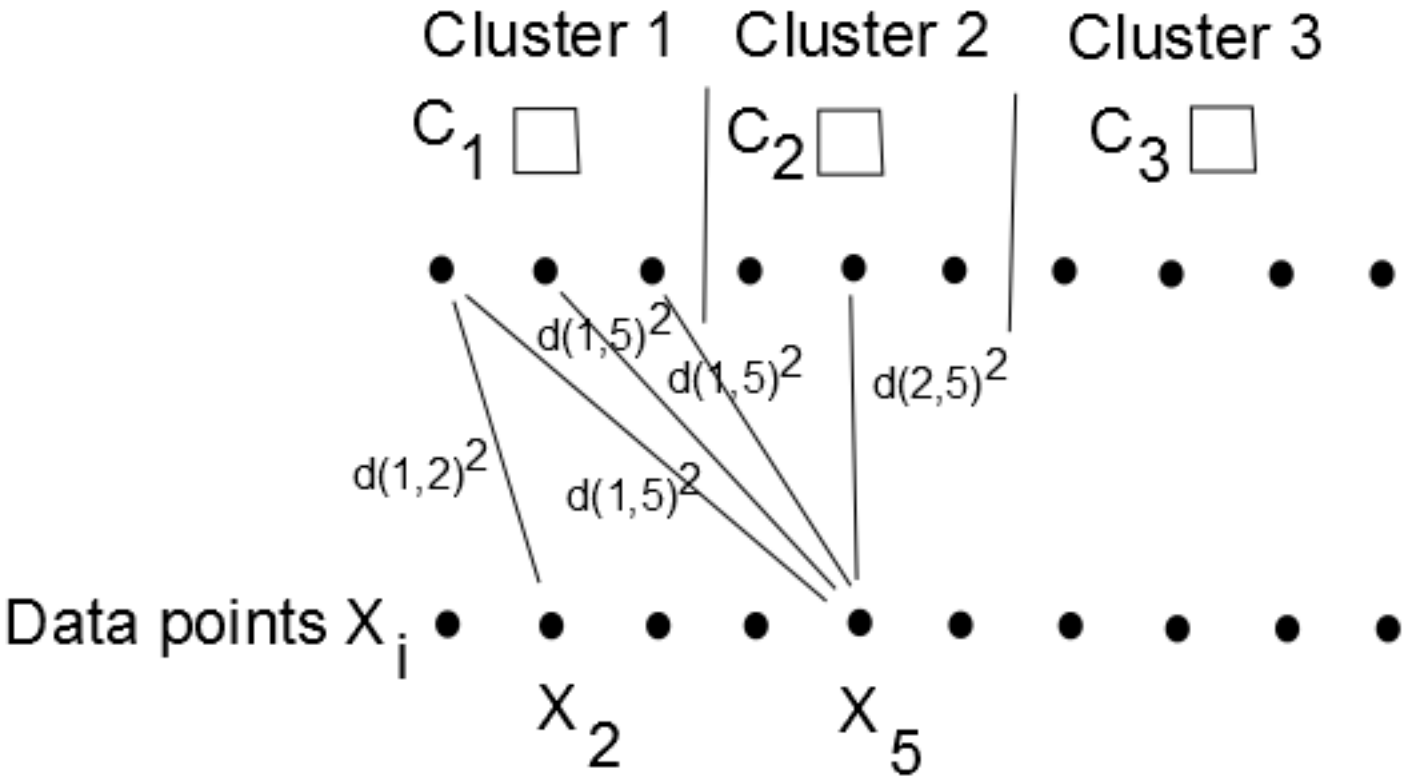
UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Forming the assignment graph



University of Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

K-means –like process

1) **Initialization** of centroid locations
2) **Assignment phase**. Calculate arclengths and use Hungarian algorithm to assign.
3) **Update phase**. Calculate new centroid locations.

Repeat phases 2)-3) until centroid locations do not change.

UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Time complexity

Balanced k-means:

$$T(n) = O(I \cdot n^3)$$

Constrained k-means (from literature):

$$T(n) = O(I \cdot k^{3.5} n^{3.5})$$

Where *I* = number of kmeans iterations

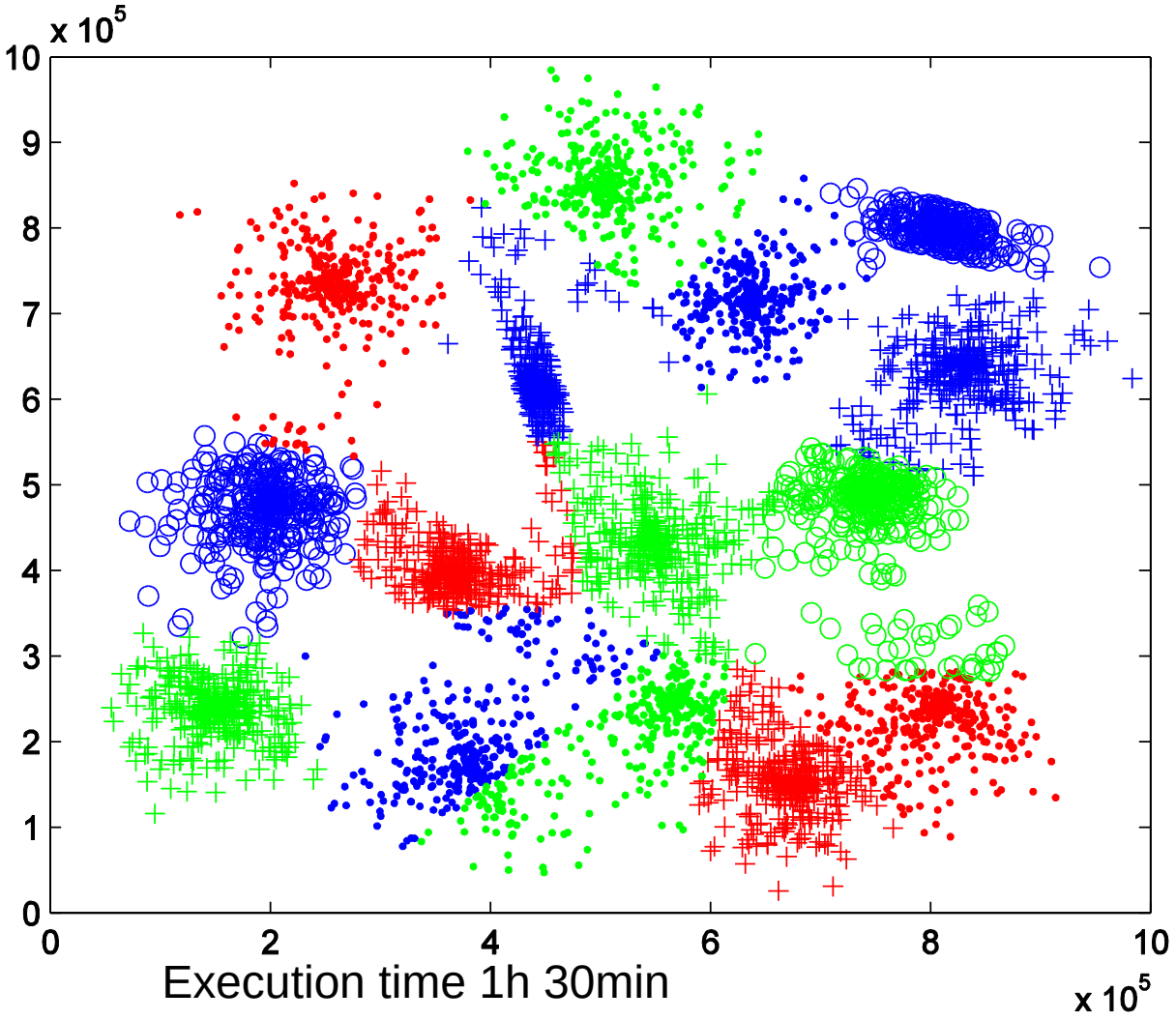
UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Experiment with dataset s2 of size n=5000

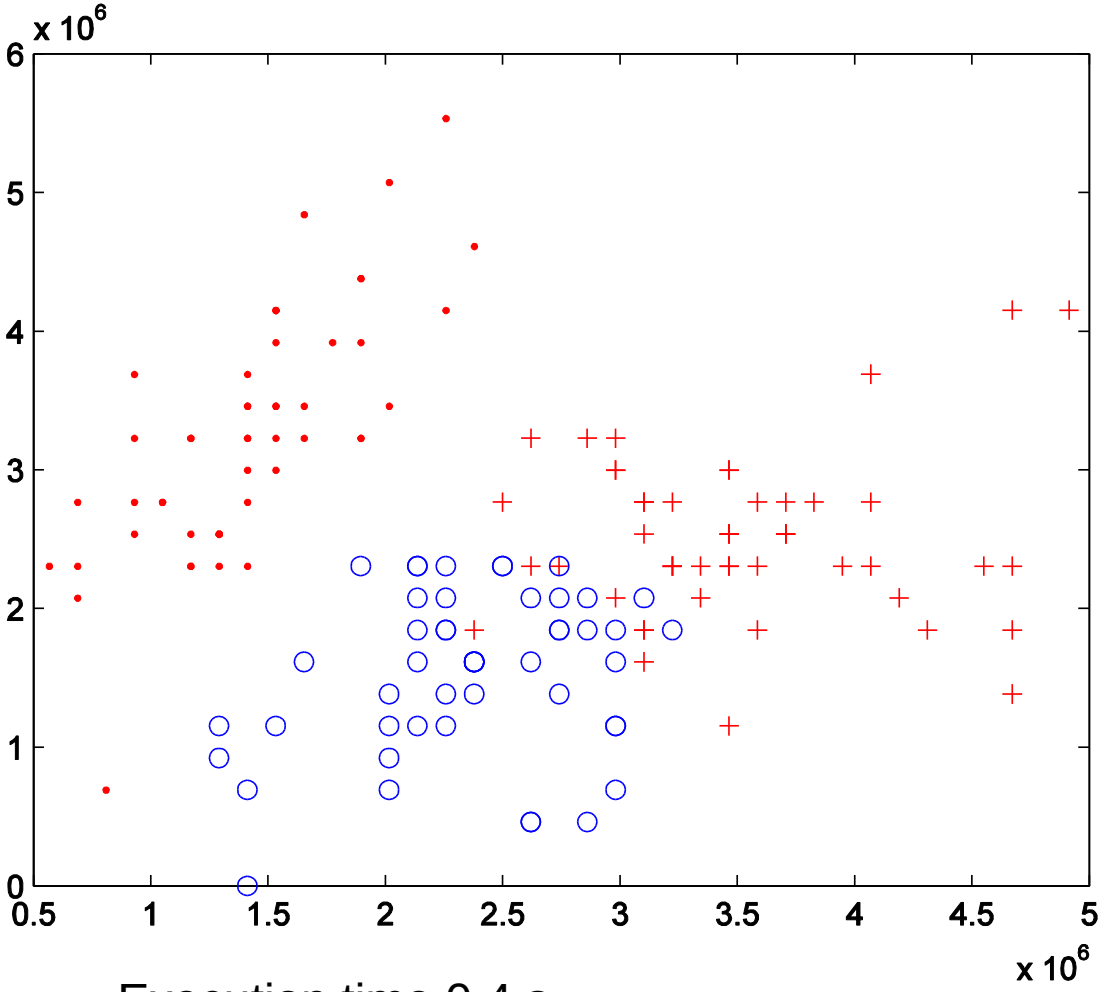


University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Experiment with Iris dataset of 150 points and 4 dimensions



Execution time 0.4 s

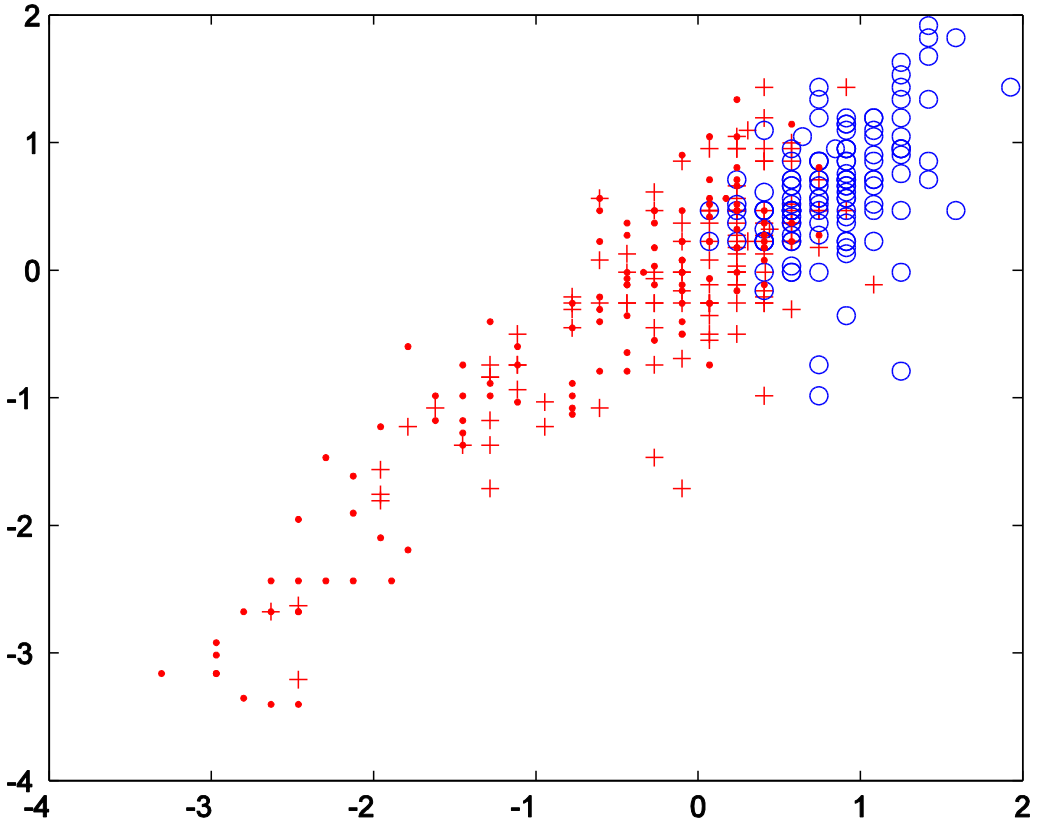
UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs

Rottavalikoitu data, normalized, dimensions 2 and 3 out of 14



Execution time 2.6 seconds

UNIVERSITY OF EASTERN FINLAND

University of
Eastern Finland
School of Computing
P.O. Box 111
FIN- 80101 Joensuu
FINLAND

Tel. +358 2944 53648
fax +358 13 251 7955

www.uef.fi/cs