Balanced Clustering with Least Square Regression

Team7 (Onkar Verma, Navdeep Singh Chahal, Ashish Kumar & Priyansh Agrawal)

Introduction

- Clustering algorithms find use in many applications.
- Problems occurs when we want clusters to be balanced. Clustering algorithms often produce unbalanced structures or trees that are difficult to navigate.
- Examples like, Energy Load Balance of wireless sensor networks where unbalanced cluster structure may cause unbalanced energy consumption and shorten the network lifetime. Similarly, in case of **photo query system**, retail chain problem etc. Balanced clusters plays an important role in order to produce best results more quickly.

Introduction (Cont.)

- In this paper, the authors proposed a novel and simple method for clustering (BCLS), to minimize least square linear regression with balance constraint to regularize the clustering model.
- The algorithm proposed is best among all state-of-art algorithms so far.

Related Work

- Over the past decades, many clustering algorithms have been proposed and extended, such as **K-means, fuzzy C-means, spectral clustering methods,** and **projected clustering**.
- Some of closely related work includes clustering algorithms that are able to produce balanced clusters. These algorithms can be categorized into two types, namely, **Hard-balanced clustering algorithms** in which 'cluster size' is strictly required by setting fixed number of samples in the clusters and **Soft-balanced clustering algorithms** in which balancing is the aim but not mandatory requirement.

Related Work (Cont.)

- Algorithms like ‘Constraint K-means’ (Bradly, Benneet, Demiriz 2000) and ‘Balanced K-means’ (Malinen and Franti, 2014) are two Hard-balanced clustering algorithms. Where as works of Banerjee and Ghosh (2002 and 2004), Zhong and Ghosh (2003) and Chang et.al (2014) are some of the evidence based on Soft-Balanced Clustering.

How current work is better than previous

- The aim of paper is to have balanced result i.e, balanced number of clusters and higher cluster quality i.e, high clustering accuracy at the same time without compromising either one.
- Consider a balance constraint to regularize the clustering model, which belongs to the soft-balanced algorithms, in order to have a balanced result and maintain good clustering performance simultaneously.

How current work is better than previous (Cont.)

- Incorporated least square linear regression as it can provide the model of dissimilarity for clustering to guide the partitioning. Also, it can estimate the class-specific hyperplanes dividing each class of data from others.
- Experimented on 7 benchmarks datasets and observed that the proposed approach not only produces good clustering performance but also guarantees balanced clustering result.

Least Square Linear Regression

- In regression, the estimation error is minimized as follows: $\min \sum_{i=1}^{n} e(f(x_i), y_i)$, where $f(x_i) = x_i^T w + b$ for 2-class dataset $\{(x_i, y_i)\}_{i=1}^{n}$. Each $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.
- In general, for dataset with c classes, Multivariate linear regression problem is given as: $\min_{W,b} \sum_{i=1}^{n} ||W^T x_i + b - y_i||_2^2 + \gamma R(W)$. Here W is the projection matrix or matrix whose columns are normals of class-specific hyperplanes. b is the bias vector and R(W) = $||W||_F^2$

Augmented Lagrange Multipliers(ALM) Method

- ALM are a series of algorithms for solving constrained optimization problems.
- Given constraint OP of following kind:
  min f(X), s.t. H(X) = 0 ,where $f : \mathbb{R}^{mxn} \rightarrow \mathbb{R}$ and $H : \mathbb{R}^{mxn} \rightarrow \mathbb{R}^{mxn}$.
  The augmented lagrangian function is defined as:
  $L(X, \wedge, \mu) = f(X) + \langle \wedge, H(X) \rangle + \frac{\mu}{2} ||H(X)||_F^2$
  , where μ is the positive scalar that gets updated after each iteration and $\wedge$ is an estimate of the Lagrange multiplier with the estimation accuracy improved at every step.

Background (Cont.)

- The constraint OP becomes an unconstraint problem by minimizing the Lagrangian function L(X,$\wedge$, μ) with updating parameters μ and $\wedge$.
- The indication of convergence of the method is H(X) $\rightarrow$ **0** or $\wedge$ remains unchanged.

Notations and Variables used in the paper

- M=$(m_{ij}) \in \mathbb{R}^{pxq}$, m_{ij} denotes the (i,j)-th entry of M, M^T denotes the transpose of M, and tr(M) denotes the trace of M. The F-norm of M is denoted by $||M||_F$ and the l_2-norm is denoted by $||M||_2$. The inner product of matrices A and B is denoted by $\langle A.B \rangle$. **1** denotes the vector with all elements as 1, and **0** denotes the vector with all elements 0.
- X = $[x_1, x_2, ..., x_n] \in \mathbb{R}^{dxn}$, where n is the number of samples and d is the data dimension. Each $x_i \in \mathbb{R}^d$. This is the mean centered data-matrix.
- Y = $[y_1, y_2, ..., y_n]^T \in \mathbb{R}^{nxc}$, is the label matrix also known as class indicator matrix ($Y \in$ Ind).

Notations and Variables used in the paper (Cont.)

- W = $[w_1, w_2, ..., w_c] \in \mathbb{R}^{dxc}$, is the projection matrix where the w_k denotes normal vector to the hyperplane that partition the k-th class from the others.
- R(W) is the regularization term of standard least square linear regression. This value equals to $||W||_F^2$.
- b $\in \mathbb{R}^c$, is the bias vector.
- γ is the regularization parameter.
- λ is the balance parameter.

Approach to the problem (Timeline)

- Finding the balance constraint which is responsible for balanced clustering.
- Incorporating balance constraint in the optimization problem of least square linear regression.
- Solving the resultant OP using Augmented Lagrange Multiplier method (ALM).
- Performing experiment on datasets and comparing the results.

Balance Constraint

- Defining s = $[s_1, s_2, ..., s_c] \in \mathbb{R}^{1xc}$, where s_i denotes the number of samples in i-th cluster. It can be observed that s=$\mathbf{1}^T$Y , where Y is the indicator matrix.
- Average number of clusters in each cluster is $\frac{n}{c}$.
- Objective is to partition the samples into balanced clusters i.e, to make the cluster size as close to $\frac{n}{C}$ as possible. This is equivalent to minimize variance (σ^2) of s_k.

Problem Formulation (Cont.)

$$
\begin{aligned}
\min_s \sigma^2 &\Leftrightarrow \min_s \frac{1}{c}\sum_{k=1}^{c}(s_k - \frac{n}{c})^2 \\
&\Leftrightarrow \min_s \sum_{k=1}^{c}(s_k^2 - 2s_k\frac{n}{c} + \frac{n^2}{c^2}) \\
&\Leftrightarrow \min_s (\sum_{k=1}^{c} s_k^2 - \frac{n^2}{c}) \\
&\Leftrightarrow \min_s \sum_{k=1}^{c} s_k^2
\end{aligned}
\tag{1}
$$

- From above deduction, we can see that

$$
\sum_{k=1}^{c} s_k^2 = ||s||_2^2 = ||\mathbf{1}^T Y||_2^2 = tr(Y^T \mathbf{1}\mathbf{1}^T Y) \tag{2}
$$

Problem Formulation (Cont.)

- This means to minimize the variance, we need to minimize the $tr(Y^T \mathbf{1}\mathbf{1}^T Y)$. Thus this value indicates the balance degree of our clustering algorithm.
- By minimizing the $tr(Y^T \mathbf{1}\mathbf{1}^T Y)$, the data samples tend to be clustered into c balanced classes with $\frac{n}{c}$ samples in each class.

Objective Function

- Traditional least square regression model:

$$\min_{W,b} ||X^T W + \mathbf{1}b^T - Y||_F^2 + \gamma ||W||_F^2 \tag{3}$$

- Including balance constraint as another regularization term in our clustering model, we proposed BCLS with the following objective function:

$$\min_{W,b,Y \in Ind} ||X^T W + \mathbf{1}b^T - Y||_F^2 + \gamma ||W||_F^2 + \lambda tr(Y^T \mathbf{1}\mathbf{1}^T Y) \tag{4}$$

Problem Formulation (Cont.)

- In the objective function in Eq.(4), the minimization of the least square regression term guides the clustering process to partition data points into c clusters. Meanwhile, minimizing the balance regularization term guarantees the balanced partitioning among different categories.

Solving OP using ALM

- Problem is NP-hard. To solve it in polynomial time we first apply Alternating Direction Method of Multipliers (ADMM) proposed in Eckstein and Bertsekas, 1992. We replace Y with Z that has entries with continuous values, so as to transfer Eq.(4) into an equality constraint OP and approximately obtain the optimal solution by alternatively solving Y with Z fixed and solving Z with Y fixed.
- Converted equation thus becomes:

$$\min_{\substack{W,b,Y\in Ind,\\ Y=Z}} ||X^TW + \mathbf{1}b^T - Y||_F^2 + \gamma||W||_F^2 + \lambda tr(Z^T\mathbf{1}\mathbf{1}^TZ) \tag{5}$$

st. Y-Z=0

Solving OP using ALM (Cont.)

- The resultant problem in Eq.(5) is non-convex and since objective function has multiple unknown variables (W,b,Y,Z). It can be solved by alternating updating the unknown variables.
- Final solutions obtained are:

$$\begin{cases} W & = (XX^T + \gamma I_d)^{-1} XY \\ b & = \frac{1}{n} Y^T \mathbf{1} \end{cases} \tag{6}$$

Solving OP using ALM (Cont.)

$$Z = (\mu I_n + 2\lambda \mathbf{1}\mathbf{1}^T)^{-1}(\mu Y + \wedge) \tag{7}$$

- Inverse of $\mu I_n + 2\lambda \mathbf{1}\mathbf{1}^T$ can be given by $\frac{(\mu+2n\lambda)I_n - 2\lambda \mathbf{1}\mathbf{1}^T}{(\mu^2+2n\lambda\mu)}$
- Y=$(y_{ik}) \in \mathbb{R}^{nxc}$ and its entries y_{ik} is given by:

$$y_{ik} = \begin{cases} 1 & \text{if } k = arg\max_k \{v_{ik}\}_{k=1}^{c} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where, v_{ik} are the entries of matrix V = $(v_{ik}) \in \mathbb{R}^{nxc}$ and V is given by:

$$V = \frac{2}{2+\mu}(X^T W + \mathbf{1}b^T) + \frac{1}{2+\mu}(\mu Z - \wedge) \tag{9}$$

- This is obtained by solving following OP:

$$\min_{Y \in Ind} ||Y - V||_F^2 + const. \tag{10}$$

Experiment

- We have done experiment on 3 datasets namely, Ionosphere, Wine and UMIST
- Optimal values are chosen for different parameters as given in original paper.
- The time complexity in a single iteration is $O(n^2c + d^2c)$

Principal Component Analysis

Dataset	Samples	Orinigal Dim.	PCA Dim.	#Class
Wine	178	13	10	3
UMIST	380	10304	50	20
Ionosphere	351	34	20	2

Experiment (Cont.)

- Clustering accuracy (ACC) [Cai, He, and Han 2005] is used to evaluate clustering performance.
- To evaluate balancing performance, Normalized entropy (N_{entro}) [Zhong and Ghosh 2003] is incorporated.
- N_{entro} of 1 means perfectly balanced clusters and 0 means extremely unbalanced clusters.
- Below given is table for each performance metrics evaluated for each of 3 datasets.

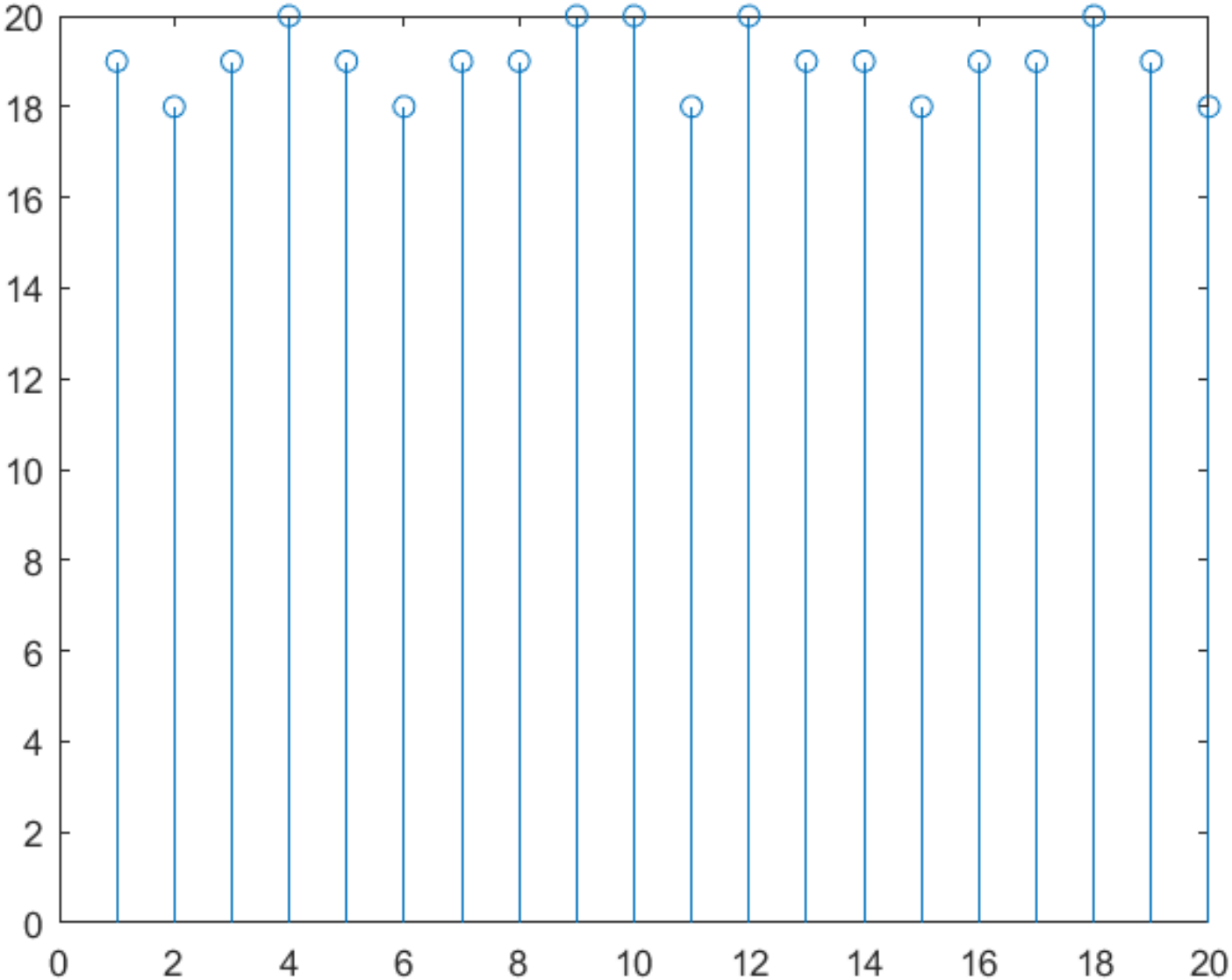
Accuracy

Dataset	correct	n	ACC	N_{entro}
Ionoshpere	239	351	68.09%	1.0
Wine	167	178	93.82%	0.9998
UMIST	224	380	61.57%	1.0

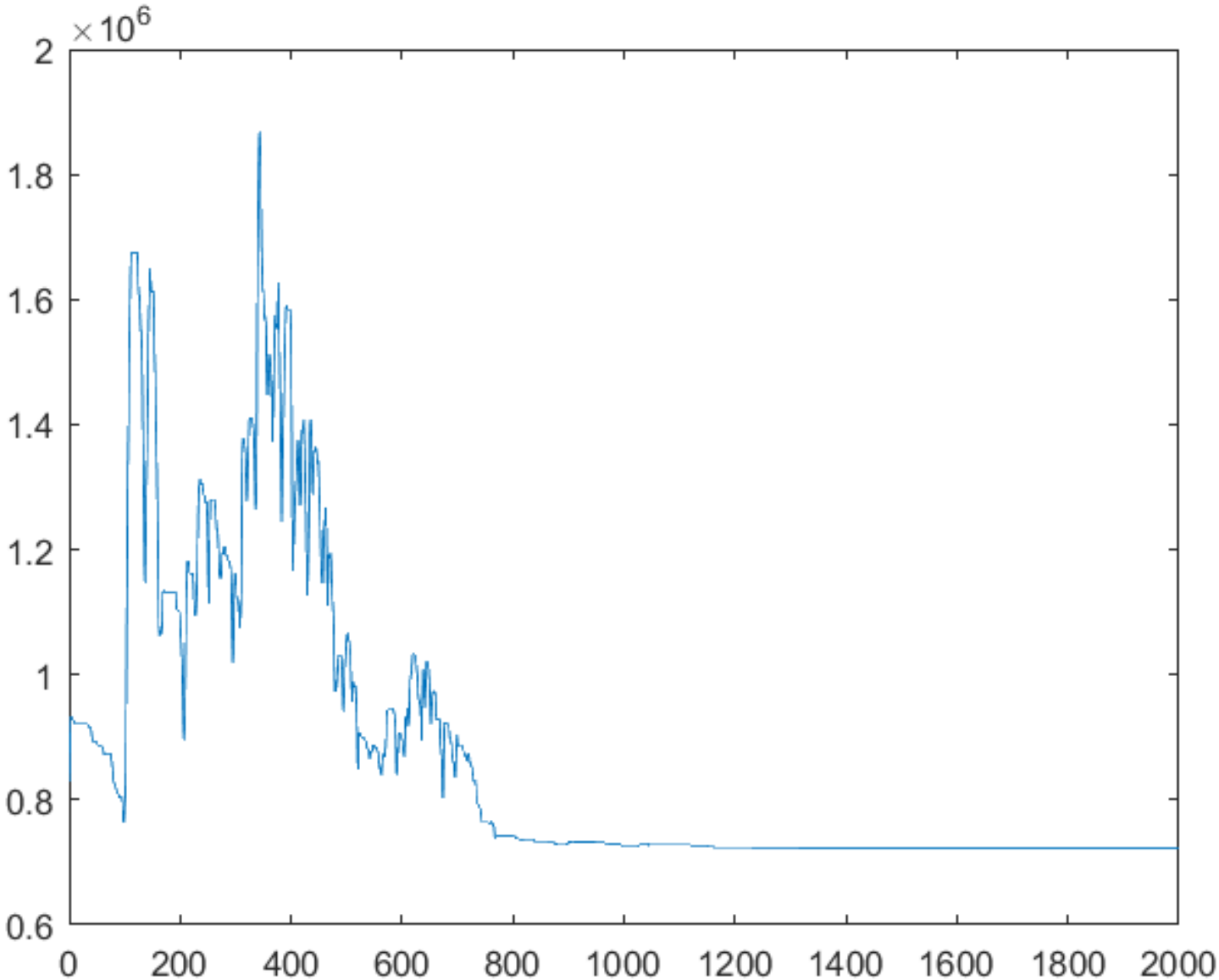
Optimal parameters

Dataset	γ	λ	μ
Ionoshpere	10^{-5}	1000	0.1
Wine	10^{-5}	10	1
UMIST	10^{-5}	100	0.1

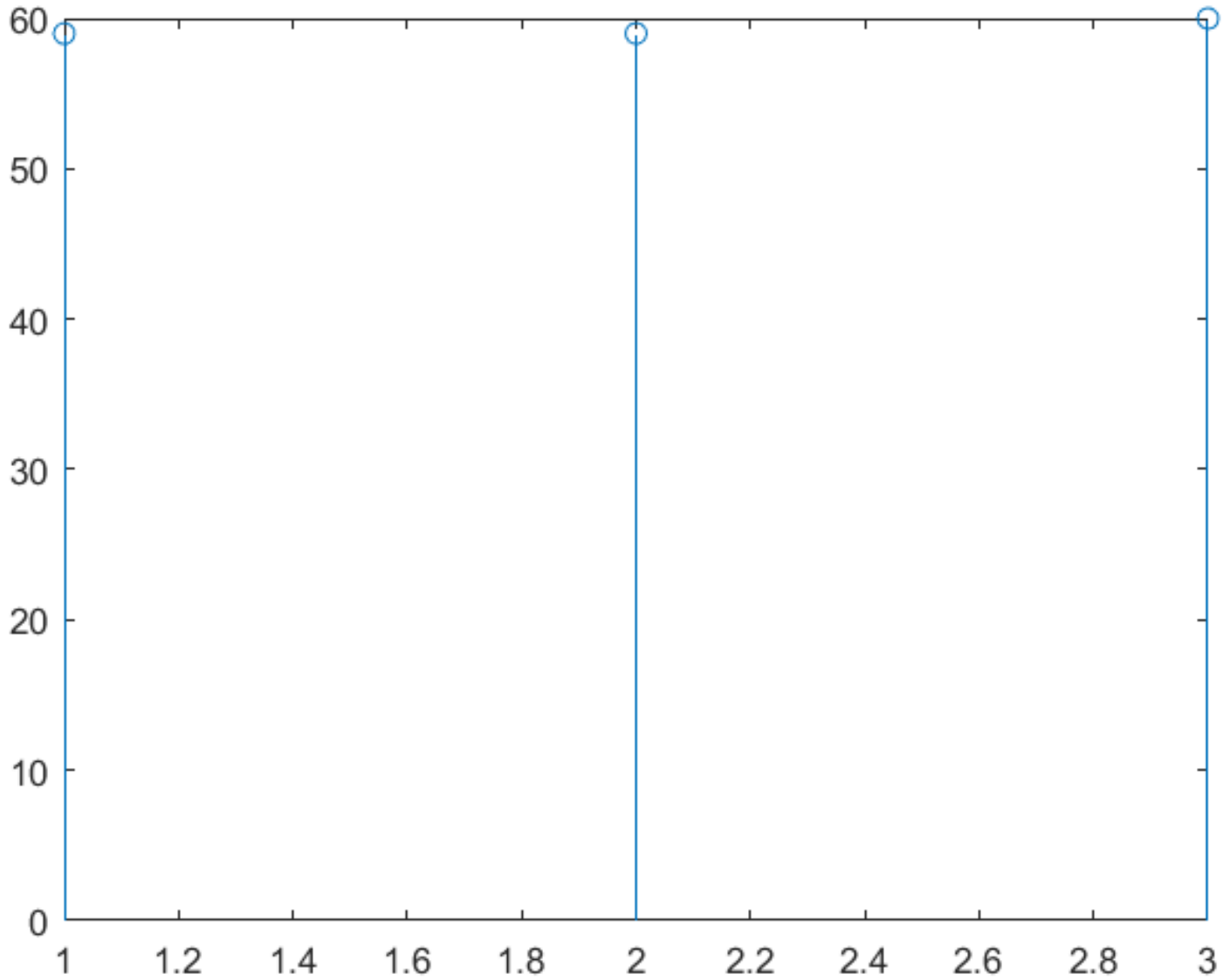
UMIST Dataset Cluster distribution



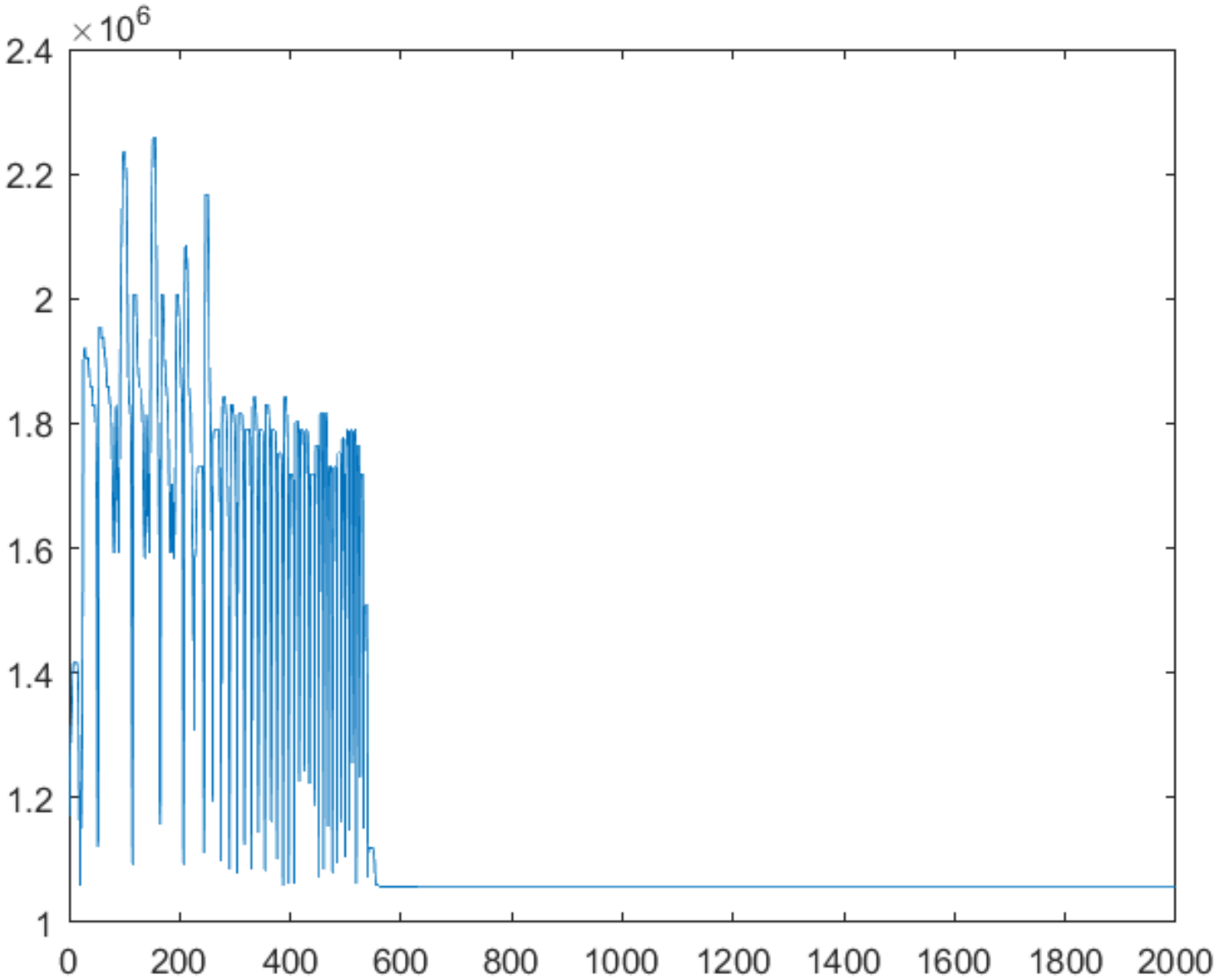
UMIST Dataset Cluster distribution



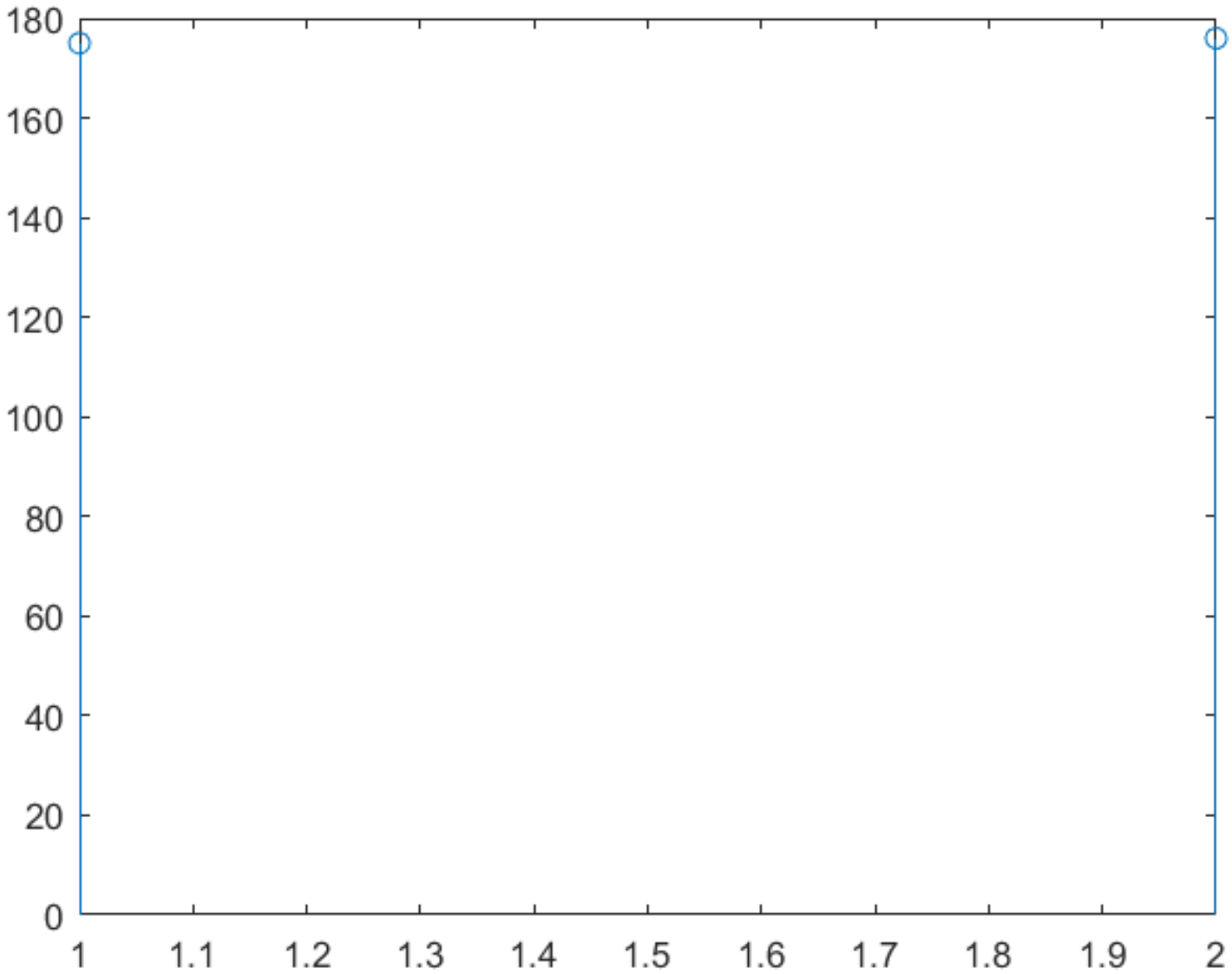
Wine Dataset Cluster distribution



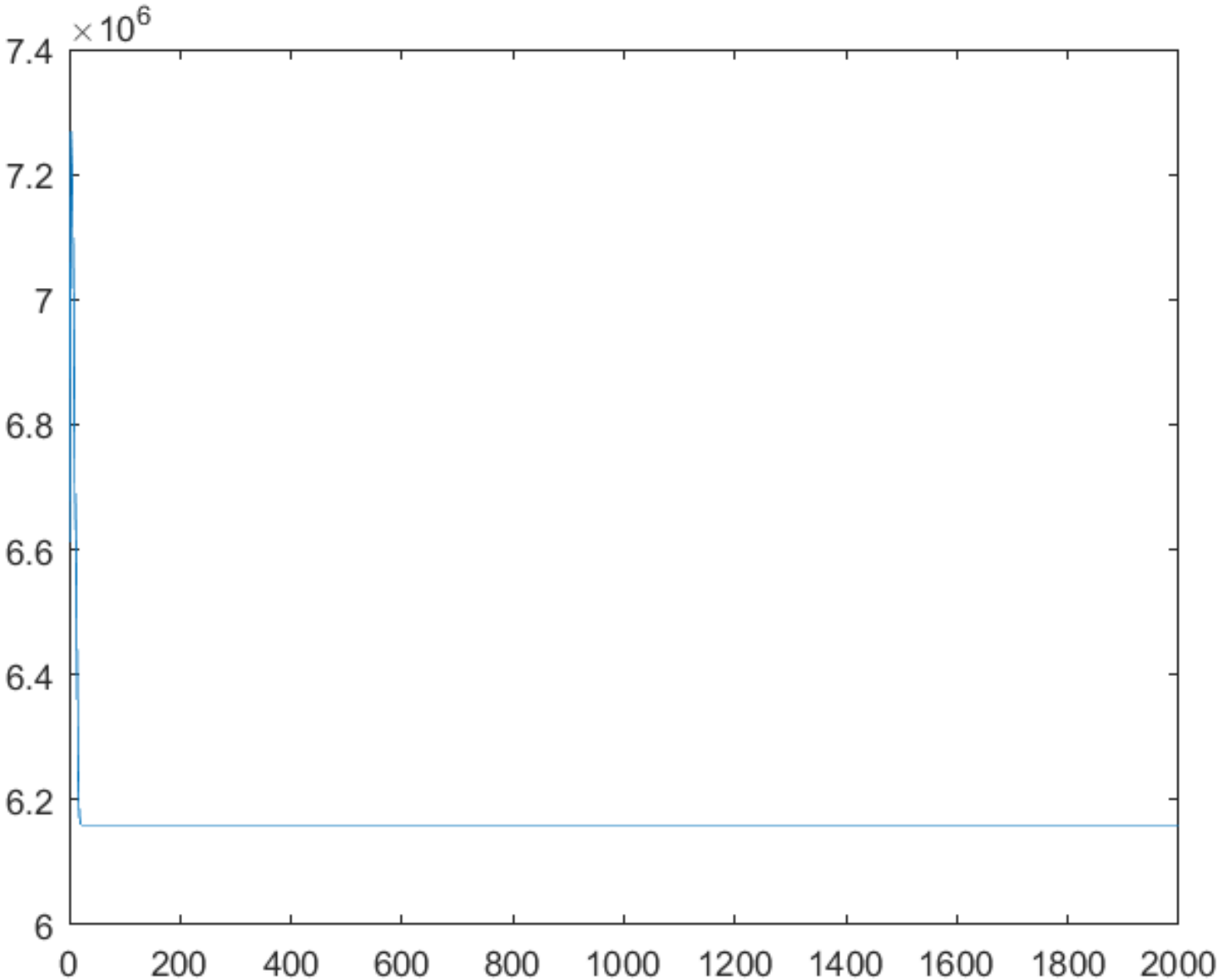
Wine Dataset Objective Function



Iono Dataset Cluster distribution



Iono Dataset Object Function



Conclusion

- Proposed a conceptually simple but effective clustering algorithm that produces balanced clusters.
- Estimated the class-specific hyperplanes that partition the data points into different clusters by iteratively minimizing the least square error of the linear regression.
- Balance constraint is introduced in order to achieve a balanced clustering result.

Conclusion (Cont.)

- ALM along with ADMM is applied for obtaining good approximate solutions of the resultant OP.
- Experiments is performed on benchmarks datasets from which we observed that BCLS produces good clustering and balancing performances simultaneously.