Modified Convex Data Clustering Algorithm Based on Alternating Direction Method of Multipliers

Tahereh Esmaeili Abharian[a*], Mohammad Bagher Menhaj[b]

[a] Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b] Department of Electrical Engineering Amirkabir University of Technology, Tehran, Iran

Abstract

Knowing the fact that the main weakness of the most standard methods including k-means and hierarchical data clustering is their sensitivity to initialization and trapping to local minima, this paper proposes a modification of convex data clustering in which there is no need to be peculiar about how to select initial values. Due to properly converting the task of optimization to an equivalent convex optimization problem, the proposed data clustering algorithm can be indeed considered as a global minimizer. In this paper, a splitting method for solving the convex clustering problem is used called as Alterneting Direction Method of Multipliers (ADMM), a simple but powerful algorithm that is well suited to convex optimization. We demonstrate the performance of the proposed algorithm on real data examples. The simulation result easily approve that the Modified Convex Data Clustering (MCDC) algorithm provides separation more than the Convex Data Clustering (CDC) algorithm. Furthermore, complexity of solving the second part of MCDC problem is reduced from $O(n^2)$ to $O(n)$.

Keywords: convex data clustering, initialization, global minimizer.

1. Introduction

Data Clustering is an important and noteworthy issue in Machine learning. Data Clustering is the task of grouping a set of points in such a way that points in the same group (called a cluster) are closer to each other than to those in other groups. Different kinds of data clustering algorithms have been proposed[1];

J. McQueen et al. proposed k-means [2]method as a method using the squared error. Fuzzy versions of methods based on the squared error were defined, beginning with the Fuzzy C-Means by James C. Bezdek et al [3].Hierarchical clustering [4-6]aims to obtain a hierarchy of clusters, called dendrogram, that shows how the clusters are related to each other. These methods proceed either by iteratively merging small clusters into larger ones (agglomerative algorithms, by far the most common) or by splitting large clusters (divisive algorithms).

Density-based methods including DBSCAN [7] which is proposed by Martin Ester et al, consider that clusters are dense sets of data items separated by less dense regions; clusters may have arbitrary shape and data items can be arbitrarily distributed. Many of the graph-theoretic clustering methods are also related to density-based clustering. The data items are

* Corresponding author. Email: tahereh.esmaili@gmail.com

represented as nodes in a graph and the dissimilarity between two items is the “length” of the edge between the corresponding nodes. In several methods, a cluster is a sub graph that remains connected after the removal of the longest edges of the graph [8]. Based on graph-theoretic clustering, there has been significant interest recently in spectral clustering using kernel methods [9]. Mixture-resolving methods assume that the data items in a cluster are drawn from one of several distributions (usually Gaussian) and attempt to estimate the parameters of all these distributions. The introduction of the expectation maximization (EM) algorithm in [10] was an important step in solving the parameter estimation problem.

To ensure optimization problems are globally converged, convexification of non-convex problems has recently attracted scientists’ attentions. Lindsten et al. [11] and Hocking et al. [12] formulate the clustering task as a convex optimization problem.

This paper proposes Modified Convex Data Clustering (MCDC)algorithm in which there is no need to be peculiar about how to select initial values. Due to properly converting the task of optimization to an equivalent convex optimization problem, the proposed data clustering algorithm can be indeed considered as a global minimizer.

The rest of the paper is organized as follows:

In part 2 Convex Data Clustering (CDC) Algorithm is reviewed. In part 3 Modified Convex Data Clustering (MCDC) Algorithm is proposed and in part 4 Alternating Direction Method of Multipliers (ADMM) Algorithm that is used to solve MCDC algorithm is reviewed. In part 5 Solving MCDC problem using ADMM is proposed. Finally, the results and conclusion are explained in the last two parts.

2. Convex Data Clustering (CDC) Algorithm

Kmeans data clustering is a common method for clustering data. Although implementation of Kmeans algorithm is very simple, due to it’s non-convex formulation, this algorithm has some weaknesses such as sensitivity to initialization, predefined number of clusters and having no guarantee to get global convergence. In [11], a convex formulation for data clustering was proposed in which there is no need to be curious about how to select initial values. In addition, the number of clusters is specified dynamically. Since, the number of clusters is not predefined, for each pattern *(x_j)*, it is considered μ_j that presents cluster’s center that x_j belongs to. Two x_j belong to one cluster if their corresponding μ_j are the same:

$$\min_x \sum_{j=1}^{N} \left\|x_j - \mu_j\right\|^2 \qquad \text{s.t. } \{\mu_1, \dots, \mu_N\} \tag{1}$$

In order to adaptively tune the number of clusters, a tuning expression is proposed. So data clustering problem is defined as the following:

$$\min_x \sum_{j=1}^{N} \left\|x_j - \mu_j\right\|^2 + \lambda \sum_{j=2}^{N} \sum_{i<j} \left\|\mu_i - \mu_j\right\|_p \tag{2}$$

$\{\boldsymbol{\mu_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{N}}$contains N vectors which k number of them are unique (k clusters). In the optimal point, for some i,j, $||\mu_i\text{-}\mu_j||_p$ is equal to zero and corresponding patterns *(x_j,x_i)* belong to same cluster. Therefore the number of clusters reduces efficiently. It makes possible to control the number of clusters by properly tuning the parameter λ. The penalty expression in convex data clustering was multiplied by a weight vector [12] in a way that, as distances between points are increased, the weight vector is reduced. In other words, for the points which are far away from each other, there is a few needs to put pressure on corresponding centers of clusters to become close to each other.

In [5] the relation between this method of data clustering and hierarchical data clustering is explained.

There are key features that make this data clustering method appropriate:

- The aforementioned optimization problem is convex. Therefore there is no need to be curious about how to select initial values. Due to properly converting the task of optimization to an equivalent convex optimization problem, the CDC algorithm can be indeed considered as a global minimizer. Lots of common data clustering methods, including Kmeans should have an appropriate initialization to gain a good result.
- In CDC algorithm, there is no need to predefine the number of clusters. In other words, the number of clusters is controlled by using the tuning parameter λ. This feature is useful when encountering "data stream" problem in which the number of clusters is changing by time [13].

In the following Modified Convex Data Clustering (MCDC) algorithm is proposed.

3. Modified Convex Data Clustering (MCDC) Algorithm

In this section the MCDC algorithm is proposed in which unnecessary terms are by passed. For example, if there are penalties for distances between μ_1 and μ_2 and between μ_2 and μ_3, it is not really necessary to put a penalty for distances between $\mu_1$and μ_3. Therefore, the modified form of (2) becomes:

$$min_\mu \sum_{j=1}^{N} \|x_j - \mu_j\|^2 + \lambda \sum_{j=1}^{N-1} \|\mu_j - \mu_{j+1}\|_p \qquad (3)$$

It should be noticed that this index helps economizing storage space noticeably. It can be easily shown that complexity of solving the second part of(3)is reduced from $O(n^2)$to $O(n)$. Because two nested loops in second part of (2) is reduced to one loop.

By adding weights (w_{jj+1}) to (3), we will have:

$$\min_\mu \sum_{j=1}^{N} \|x_j - \mu_j\|^2 + \lambda \sum_{j=1}^{N-1} w_{jj+1} \|\mu_j - \mu_{j+1}\|_p \qquad (4)$$

Furthermore, ifl_1 Norm is used, the underlying computation can be remarkably reduced. w_{jj+1} could be presented by:

$$w_{jj+1} = e^{-\varphi\|x_j - x_{j+1}\|_2^2} \qquad (5)$$

The coefficient φ is a scalar value where 0<φ<1.

It is important to note that, if the CDC algorithm is used instead of MCDC algorithm, the distances between "each" two points should be computed for measuring weights.

In the following section, Alternating Direction Method of Multipliers (ADMM) Algorithm for solving (4)is explained.

4. Alternating Direction Method of Multipliers Algorithm

ADMM is an algorithm to solve problems given in (6):

$$minimize f(x) + g(z) \qquad (6)$$

$$subject\ to\ Ax + Bz = c$$

With variables $x \in R^n$ and $z \in R^m$, where $A \in R^{p\times n}$, $B \in R^{p\times m}$ and $c \in R^p$. This algorithm assumes that functions f and g to be closed, proper, and convex. The optimal value of the problem (6) will be denoted by:

$$p^* = inf_{x,z}\{f(x) + g(z) | Ax + Bz = c\} \qquad (7)$$

As in the method of multipliers[14], we form the augmented Lagrangian:

$$L_\rho(x,y,z) = f(x) + g(z) + y^T(Ax + Bz - c) + (^\rho/_2)\|Ax + Bz - c\|_2^2 \qquad (8)$$

ADMM algorithm summarized as follows:

$$x^{k+1} := argmin_x L_p(x, z^k, y^k)$$

$$z^{k+1} := argmin_z L_p(x^{k+1}, z, y^k) \qquad (9)$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

Many variations on the classic ADMM algorithm have been explored in the literature [15-18]. Below we present the convergence issue of ADMM algorithm.

4.1 Convergence

There are many convergence results for ADMM discussed in the literature [19, 20]; here, we limit ourselves to a basic but still very general result. We will make one assumption about the functions f and g, and one assumption about problem (6).

Assumption 1: The (extended-real-valued) functions f:$R^n \rightarrow R\cup\{+\infty\}$ and g:$R^n \rightarrow R\cup\{+\infty\}$are closed, proper, and convex.

This assumption can be expressed compactly using the epigraphs of the functions: The function f satisfies assumption 1 if and only if its epigraph is a closed nonempty convex set.

$$epif = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} | f(x) \le t\} \qquad (10)$$

Assumption 1. implies that the sub problems arising in the x-update and z-update are solvable, i.e., there exist x and z, not necessarily unique (without further assumptions on *A* and *B*), that minimize the augmented Lagrangian. It is important to note that assumption 1 allows f and g to be non-differentiable and to assume the value +∞.

Assumption 2. The unaugmented Lagrangian L_0 has a saddle point.

Explicitly, there exist (x^*,z^*,y^*), not necessarily unique, for which

$$L_0(x^*,z^*,y\ \ ) \le L_0(x^*,z^*,y^*) L_0(x\ \ ,z\ \ ,y^*) \qquad (11)$$

holds for all x, z, y.

Byassumption1, it follows that $L_0\ (x^*,z^*,y^*)$is finite for any saddle point $(x^*,\ z^*,y^*)$. This implies that $(x^*,z^*)$is a solution to (6), so $Ax^*+Bz^*=c$ and $f(x^*)<\infty$, $g(z^*\ )<\infty$. It also implies that $y^*$is dual optimal, and the optimal values of the primal and dual problems are equal, i.e., that strong duality holds. Note that we make no assumptions about A, B, or c, except implicitly through assumption 2; in particular, neither A nor B is required to be full rank.

Under assumptions 1 and 2, the ADMM iterates satisfy the following:

Residual convergence:$r^k \rightarrow 0$ as $k \rightarrow \infty$i.e., the iterates approach feasibility.

Objective convergence: $f(x^k)+g(z^k\ ) \rightarrow p^*$ as $k \rightarrow \infty$, i.e., the objective function of the iterates approaches the optimal value.

Dual variable convergence: $y^k \rightarrow y^*$ as $k \rightarrow \infty$, where $y^*$is a dual optimal point.

4.2 Optimality Conditions and Stopping Criterion

The necessary and sufficient optimality conditions for the ADMM problem (6) are primal feasibility,

$$Ax^* + Bz^* - c = 0 \qquad (12)$$

and dual feasibility:

$$0 \in \partial f(x^*) + A^T y^* \qquad (13)$$

$$0 \in \partial g(z^*) + B^T y^* \qquad (14)$$

Here, ∂ denotes the sub differential operator. (When f and g are differentiable, the sub differentials ∂f and ∂g can

be replaced by the gradients ∇f and ∇g, and ∈ can be replaced by =.)

Since $z^{(k+1)}$, minimizes L_ρ $(x^{k+1},z,y^k)$by definition, we have that

$$\begin{aligned} 0 &\in \partial g(z^{k+1}) + B^T y^k + \rho B^T (Ax^{k+1} + Bz^{k+1} - c) \\ &= \partial g(z^{k+1}) + B^T y^k + \rho B^T r^{k+1} \qquad (15) \\ &= \partial g(z^{k+1}) + B^T y^{k+1} \end{aligned}$$

This means that $z^{(k+1)}$and $y^{(k+1)}$always satisfy (14), so attaining optimality comes down to satisfying (12) and (13). Since $x^{(k+1)}$minimizes L_p (x,z^k,y^k) by definition, we have that

$$\begin{aligned} 0 &\in \partial f(x^{k+1}) + A^T y^k + \rho A^T (Ax^{k+1} + Bz^k - c) \\ &= \partial f(x^{k+1}) + A^T \left(y^k + \rho r^{k+1} + \rho B(z^k - z^{k+1})\right) \end{aligned} \qquad (16)$$

$$= \partial f(x^{k+1}) + A^T y^{k+1} + \rho A^T B(z^k - z^{k+1}),$$

or equivalently,

$$\rho A^T B(z^{k+1} - z^k) \in \partial f(x^{k+1}) + A^T y^{k+1} \quad (17)$$

This means that the quantity

$$s^{k+1} = \rho A^T B(z^{k+1} - z^k) \quad (18)$$

can be viewed as a residual for the dual feasibility condition (13).We will refer to $s^{(k+1)}$as the dual residual at iteration $k + 1$, and to$r^{k+1}=Ax^{k+1}+Bz^{k+1}-c$*as* the primal residual at iteration $k + 1$.In summary, the optimality conditions for the ADMM problem consist of three conditions, (12–14). The last condition (14) always holds for $(x^{k+1},z^{k+1},y^{k+1})$. the residuals for the other two, (12) and (13), are the primal and dual residuals r^{k+1} and s^{k+1}, respectively. These two residuals converge to zero as ADMM proceeds. Convergence proof and stopping criteria are shown in [14].

5. Solving Modified Convex Data Clustering Problem Using ADMM Algorithm

At first, another form of problem (4) is presented:

$$minimize \frac{1}{2}\sum_{i=1}^{p} \|x_i - u_i\|_2^2 + \gamma \sum_l w_l \|v_l\| \quad (19)$$

$$\mathrm{s.t} \qquad \mathrm{u}_{\mathrm{l}_1} - \mathrm{u}_{\mathrm{l}_2} - \mathrm{v}_{\mathrm{l}} = 0$$

$l_2=l_1+1$ and p is representing the number of patterns (points). Augmented Lagrangian for problem (19) is:

$$\begin{aligned} l_v(U,V,\Lambda) = \frac{1}{2}\sum_{i=1}^{p} \|x_i - u_i\|_2^2 + \gamma \sum_l w_l \|v_l\| \\ + \sum_l < \lambda_l, v_l - u_{l_1} + u_{l_2} \\ > + \frac{\rho}{2}\sum_l \left\|v_l - u_{l_1} + u_{l_2}\right\|_2^2 \end{aligned} \quad (20)$$

Derivation of $l_v$with respect to u_i is:

$$\begin{aligned} \frac{\partial}{\partial u_i} l_v(U,V,\Lambda) \\ = u_i - x_i \\ - \sum_{l_1=i} \lambda_l \\ + \sum_{l_2=i} \lambda_l - \rho \sum_{l_1=i} (v_l - u_i \\ + u_{l_2}) \\ + \rho \sum_{l_2=i} (v_l - u_{l_1} + u_i) \end{aligned} \quad (21)$$

For updating U, $\partial/(\partial u_i\ )\ l_v\ (U,V,\Lambda)$ should be zero. Therefore equation (22) is attained:

$$\begin{aligned} u_i = \frac{1}{(1+2(p-1)\rho)}(x_i \\ + \rho(v_{i,i+1} - v_{i-1,i} + u_{i+1} \\ + u_{i-1}) + \lambda_{i,i+1} - \lambda_{i-1,i}) \end{aligned} \quad (22)$$

Due to $u_{i+1}+u_{i-1} \leq x_{i+1}+x_{i-1} \cong 2\bar{x}$, $2\bar{x}$is replaced with $u_{i+1}+u_{i-1}$ in equation (22).

For updating V, equation (23) is computed:

$$v_l = argmin_v \frac{1}{2}\left\|v - (u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l)\right\|_2^2 + \frac{\gamma w_l}{\rho}\|v\| \quad (23)$$

By considering $\sigma_l=\gamma w_l$ and equations (24),(25) and (26) , v_l is attained:

$$prox_{\sigma\Omega}(u) = argmin_v\ [\sigma\Omega(v) + \frac{1}{2}\|u - v\|_2^2 \quad (24)$$

$$v_l = prox_{\frac{\sigma_l\|.\|}{\rho}}(u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l) \quad (25)$$

$$v_l = S_{\frac{\gamma w_l}{\rho}}(\mathrm{u}_{\mathrm{l}_1} - \mathrm{u}_{\mathrm{l}_2} - \rho^{-1}\lambda_{\mathrm{l}}) = \quad (26)$$

$$\begin{cases} u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l - \frac{\gamma w_l}{\rho} & if u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l > \frac{\gamma w_l}{\rho} \\ 0 & \text{if } \left|u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l\right| \leq \frac{\gamma w_l}{\rho} \\ u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l + \frac{\gamma w_l}{\rho} & \text{o.w } u_{l_1} - u_{l_2} - \rho^{-1}\lambda_l < -\frac{\gamma w_l}{\rho} \end{cases}$$

And finally, Lagrangian multiplier is computed by using equation (27):

$$\lambda_l = \lambda_l + \rho(v_l - u_{l_1} + u_{l_2}) \quad (20)$$

5.1. A Modified Convexdata Clustering Algorithm Based on ADMM

```
Initialize $\lambda^0$ and $v^0$
For $m = 1,2, \dots$ do
For $i = 1,2, \dots, p$ do
        $y_i = x_i + \rho(v_{i,i+1} - v_{i-1,i} + 2\bar{x}) + \lambda_{i,i+1} - \lambda_{i-1,i}$
end For
        $U^m = \frac{1}{(1 + 2\rho(p-1))} Y$
For $i = 1,2, \dots, p - 1$ do
        $v_{i,i+1}{}^m = prox_{\gamma w_{i,i+1}\|.\|}(u_i{}^m - u_{i+1}{}^m - \rho^{-1}\lambda_{i,i+1}{}^{m-1})$
```

6. Datasets

The main dataset, which is used in this paper, contained teeth features of different kinds of mammals. In this dataset, there are 27 patterns of mammals along with 8 features which present different kinds of teeth. The main reason of selecting this data set is comparing the results with CDC algorithm [12]. Iris dataset [21] is another one which is used that contains data about three types of iris plant. This dataset consists of 150 patterns and 5 features represent appearance of flower. Another dataset that is used in this paper is mouse gene dataset [22]. It includes biological data from mouse genes and consists of 150 genius patterns along with 8 features provide cell data.

7. Results

The following results are generated based on the R programming language implementation in Linux operating system with two core processor and four gigabyte RAM.

For internal validation, we selected measures that reflect the compactness, connectedness, and separation of the cluster partitions. Connectedness relates to what extent observations are placed in the same cluster as their nearest neighbors in the data space, and is here measured by the connectivity. Compactness assesses cluster homogeneity, usually by looking at the intra-cluster variance, while separation quantise the degree of separation between clusters (usually by measuring the distance between cluster centroids). Since compactness and separation demonstrate opposing trends (compactness increases with the number of clusters but separation decreases), popular methods combine the two measures into a single score. The Dunn Index and Silhouette Width are both examples of non-linear combinations of the compactness and separation. For a good overview of internal measures in general see [22]. In Tables 1, 2, 3, the yellow cells show the best values for parameter of data clustering validity. The best value for connectivity index is the minimum one and for indices Dunn and Silhaute, is maximum one. As it is shown in the tables, MCDC algorithm provides more compactness and separation than CDC algorithm.

In Fig. 1 and Fig. 2, 27 patterns of mammals based on two principal components are shown. For each pattern there is a line between it's corresponding first cluster and it's final cluster. At first, the first data cluster for each pattern is itself and finally it converges to red points. The Fig. 1, Fig. 4 (left), Fig. 6 (left) show data clustering after modification and the results of the CDC algorithm are shown in Fig. 2, Fig. 4 (right) and Fig. 6 (right). The red points represent data cluster centers. (It is an important point that if the run of algorithm iterates more, the red points in both sides of Fig. 1, Fig. 6 converge to two points. The reason of stopping the algorithm early is, showing becoming the data clusters close together). As it is shown clearly, MCDC algorithm provides clustering with better separation than CDC algorithm. The Fig. 8, Fig. 9 show first and second error of MCDC Algorithm and CDC algorithm. The charts are representing the lower error of MCDC algorithm than CDC algorithm. For more information about measuring the first and second error refer to [14].

8. Conclusion

In this paper the Modified Convex Data Clustering algorithm is proposed and it is shown that the algorithm provides separation more than Convex Data Clustering algorithm. The charts are representing the lower error of MCDC algorithm than CDC algorithm. Furthermore, the Modified Convex Data Clustering problem is solved using Alternating Direction Method of Multipliers. Since the ADMM algorithm is well suited to distributed convex optimization, and in particular to large-scale problems arising in statistics, machine learning and related areas, we are going to propose distributed Convex Data Clustering in near future.

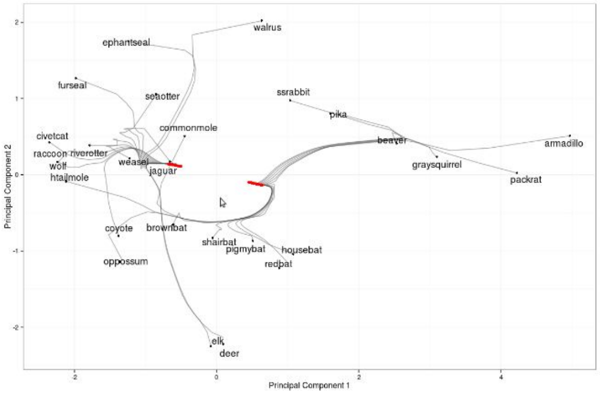

Fig. 1. Modified Convex Data clustering (mammals dataset)

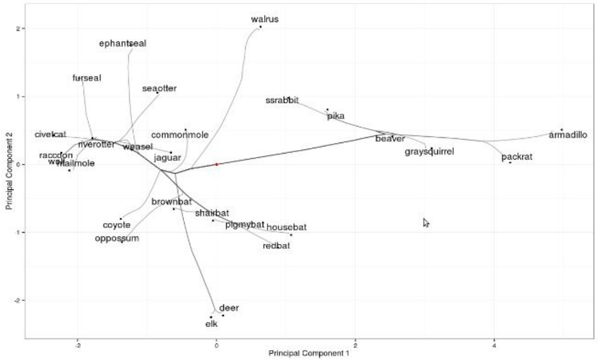

Fig. 2. Convex Data Clustering (mammals dataset)

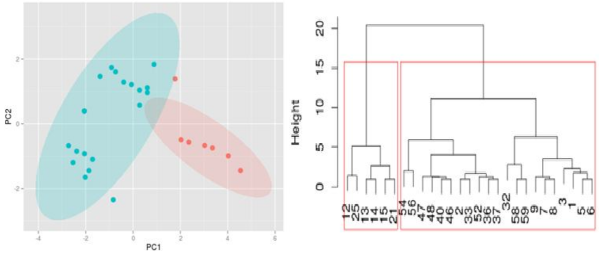

Fig. 3. Data Clustering using Kmeans (left) and hierarchical data clustering (right) (mammals dataset and the number of clusters=2)

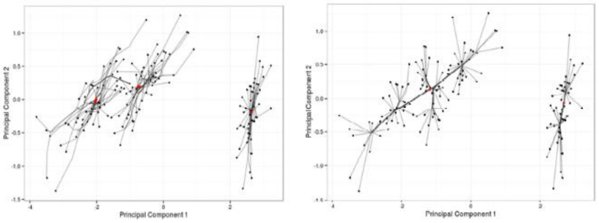

Fig. 4. Convex Data Clustering (right-two clusters) and Modified Convex Data Clustering (left-three clusters) (iris dataset)

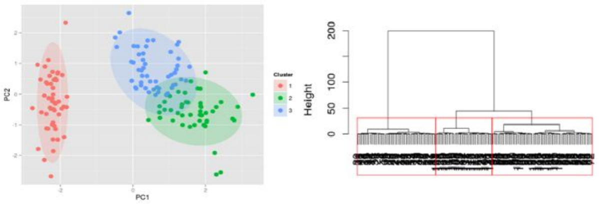

Fig. 5. Data Clustering using Kmeans (left) and hierarchical data clustering (right) (iris dataset and the number of clusters=3)

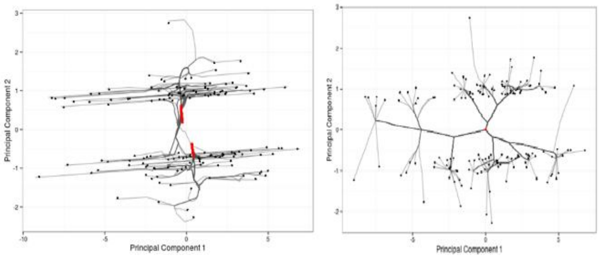

Fig. 6. Convex Data Clustering (right-one cluster) and Modified Convex Data Clustering (left-two clusters) (mousegene dataset)

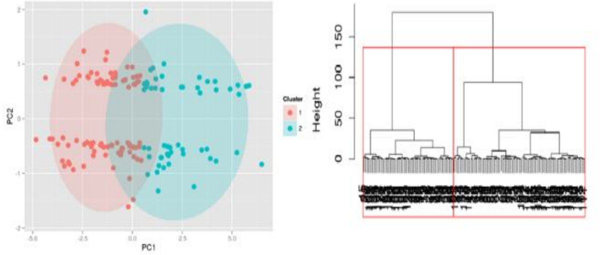

Fig. 7. Data Clustering using Kmeans (left) and hierarchical data clustering (right) (mousegene dataset and the number of clusters=2)

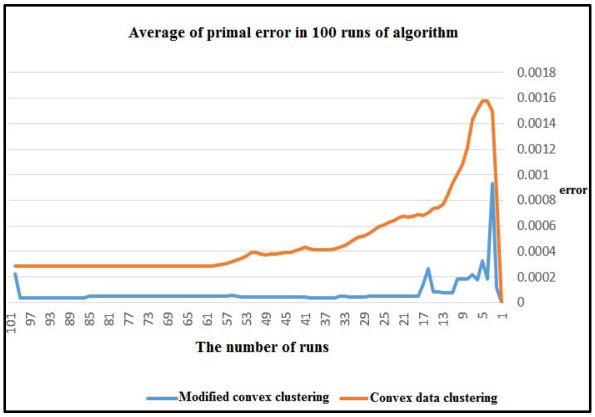

Fig. 8. Comparing primal error of Convex Data Clustering (orange) and Modified Convex Data Clustering (blue)- mammals dataset.

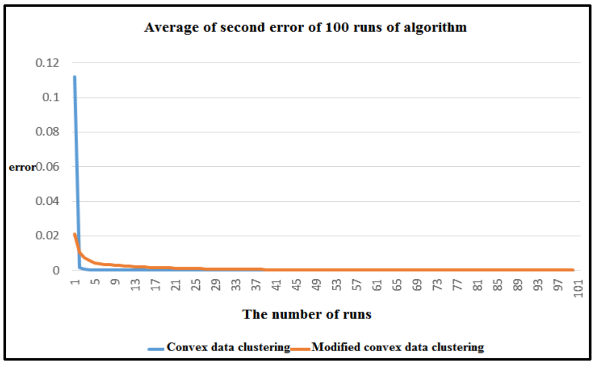

Fig. 9. Comparing second error of Modified Convex Data Clustering (orange) and Convex Data Clustering (blue)- mammals dataset.

Table 1
Cluster validity with runtime and number of clusters for mammals dataset

Iris Dataset					
	Number of clusters	Time	Connectivity	Dunn	Sillhautte
Hierarchical (ward)	3	0.052	4.4770	0.1378	0.5542
Kmeans	3	0.024	10.0917	0.0988	0.5528
Convex Data Clustering	2	1.77	0	0.3458	0.6808
Modified Convex Data Clustering	3	0.56	4.4002	0.1869	0.5511

Table 2
Cluster validity with runtime and number of clusters for iris dataset

Mousegenes Dataset					
	Number of clusters	Time	Connectivity	Dunn	Sillhautte
Hierarchical (ward)	2	0.048	5.3270	0.1291	0.3962
Kmeans	2	0.060	13.2548	0.0464	0.3911
Convex Data Clustering	1	1.3	0	0	-1
Modified Convex Data Clustering	2	0.69	5.3002	0.1314	0.5004

References

[1] R. Xu, D. Wunsch, "Survey of clustering algorithms," Neural Networks, IEEE Transactions on, vol. 16, pp. 645-678, 2005.

[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967, pp. 281-297.

[3] J. C. Bezdek, R. Ehrlich, W. Full, "FCM: The fuzzy< i> c</i>-means clustering algorithm," Computers & Geosciences, vol. 10, pp. 191-203, 1984.

[4] B. King, "Step-wise clustering procedures," Journal of the American Statistical Association, vol. 62, pp. 86-101, 1967.

[5] T. D. Hocking, A. Joulin, F. Bach, and J.P. Vert, "Clusterpath an algorithm for clustering using convex fusion penalties," in 28th international conference on machine learning, 2011.

[6] P. H. Sneath, R. R. Sokal, Numerical taxonomy. The principles and practice of numerical classification, 1973.

[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Kdd, 1996, pp. 226-231.

[8] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," in DMKD, 1997, pp. 0-.

[9] A. Y. Ng, M. I. Jordan, Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in neural information processing systems, vol. 2, pp. 849-856, 2002.

[10] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the royal statistical society. Series B (methodological), pp. 1-38, 1977.

[11] F. Lindsten, H. Ohlsson, L. Ljung, "Just relax and come clustering!: A convexification of k-means clustering," 2011.

[12] E. C. Chi and K. Lange, "Splitting Methods for Convex Clustering," arXiv preprint arXiv:1304.0499, 2013.

[13] F. Lindsten, H. Ohlsson, L. Ljung, "Clustering using sum-of-norms regularization: With application to particle filter output computation," in Statistical Signal Processing Workshop (SSP), 2011 IEEE, 2011, pp. 201-204.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends® in Machine Learning, vol. 3, pp. 1-122, 2011.

[15] Y. Ouyang, Y. Chen, G. Lan, E. Pasiliao Jr, "An accelerated linearized alternating direction method of multipliers," SIAM Journal on Imaging Sciences, vol. 8, pp. 644-681, 2015.

[16] M. Kadkhodaie, K. Christakopoulou, M. Sanjabi, A. Banerjee, "Accelerated alternating direction method of multipliers," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 497-506.

[17] E. Ghadimi, A. Teixeira, I. Shames, M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems," Automatic Control, IEEE Transactions on, vol. 60, pp. 644-658, 2015.

[18] Y. Jiao, Q. Jin, X. Lu, W. Wang, "Alternating Direction Method of Multipliers for Linear Inverse Problems," arXiv preprint arXiv:1601.02773, 2016.

[19] M. Hong, Z.-Q. Luo, M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015, pp. 3836-3840.

[20] Y. Cui, X. Li, D. Sun, K.-C. Toh, "On the Convergence Properties of a Majorized Alternating Direction Method of Multipliers for Linearly Constrained Convex Optimization Problems with Coupled Objective Functions," Journal of Optimization Theory and Applications, pp. 1-29, 2016.

[21] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of eugenics, vol. 7, pp. 179-188, 1936.

[22] G. Brock, V. Pihur, S. Datta, S. Datta, "clValid, an R package for cluster validation," Journal of Statistical Software (Brock et al., March 2008), 2011.