

RESEARCH ARTICLE

Elastic K-means using posterior probability

Aihua Zheng^{1*}, Bo Jiang¹, Yan Li², Xuehan Zhang¹, Chris Ding¹

1 Anhui University, Hefei, China, **2** Anhui Broadcasting Movie and Television College, Hefei, China

* ahzheng214@ahu.edu.cn

Abstract

The widely used K-means clustering is a hard clustering algorithm. Here we propose a Elastic K-means clustering model (EKM) using posterior probability with soft capability where each data point can belong to multiple clusters fractionally and show the benefit of proposed Elastic K-means. Furthermore, in many applications, besides vector attributes information, pairwise relations (graph information) are also available. Thus we integrate EKM with Normalized Cut graph clustering into a single clustering formulation. Finally, we provide several useful matrix inequalities which are useful for matrix formulations of learning models. Based on these results, we prove the correctness and the convergence of EKM algorithms. Experimental results on six benchmark datasets demonstrate the effectiveness of proposed EKM and its integrated model.



OPEN ACCESS

Citation: Zheng A, Jiang B, Li Y, Zhang X, Ding C (2017) Elastic K-means using posterior probability. PLoS ONE 12(12): e0188252. <https://doi.org/10.1371/journal.pone.0188252>

Editor: Feiping Nie, Northwestern Polytechnical University, UNITED STATES

Received: July 25, 2017

Accepted: November 5, 2017

Published: December 14, 2017

Copyright: © 2017 Zheng et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The AT&T dataset is available at: AT&T: <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>; The USPS, MNIST and BinAlpha datasets are available at: <http://www.cs.nyu.edu/~roweis/data.html>; The COIL-20 dataset is available at: <http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>; The Isolet1 Dataset is available at: <http://archive.ics.uci.edu/ml/datasets/ISOLET>.

Funding: This work was funded by the Natural Science Foundation of Anhui Province (1508085QF127) and the Natural Science

Introduction

Data clustering, a method of unsupervised learning and a common technique for statistical data analysis has currently been widely used in machine learning [1, 2], pattern recognition [3, 4], image analysis [5, 6] and bioinformatics [7–9]. As one of the most popular clustering methods, K-means has drawn lots of attention and been widely used for data clustering. Recent work has revealed that K-means can be represented by matrix factorization formulation, and thus can deduce many kind of K-means variants [10–15]. From the point of view of determinacy, K-means clustering approaches are hard clustering where a data point exactly belongs to one particular class. However, in many cases, this is not realistic since not all the data point distinctly belongs to one single class especially for the outliers which may have equivalent similarities to every/some classes.

On the other hand, above K-means inspired methods generally deal with the attribute information (feature vectors) of data. In real world applications, besides attribute information, we have various pairwise relations between data points which are expressed as **graph** data. There exist many clustering methods that utilize graph data, such as Ratio Cut based methods [16, 17] and Normalized Cut based methods [18, 19] and Min and Max Cut based methods [20, 21]. In essence, these clustering methods first embed graph nodes in low-dimensional space using linear embedding method PCA and nonlinear method IsoMAP [22–24], Local linear Embedding (LLE) [25–27], Local Tangent Space Alignment [26, 28, 29] etc, where feature vector information is utilized to obtain the final clustering results. However, It's one-sided to just take into account of one single information.

Foundation of Anhui Higher Education Institutions of China (KJ2016A114).

Competing interests: The authors have declared that no competing interests exist.

Inspired by recent works on matrix factorization based K-means models, In this paper, we first propose a **Elastic** clustering model using posterior probability (named as Elastic K-means, EKM). The main observation of our Elastic K-means model is that the clustering indicator from K-means can be interpreted as the posterior cluster probabilities during the factorization procedure. The key point of the Elastic K-means is that each data point is assigned to the clusters scattered into several possible classes according to its posterior probabilities. Then we extend our EKM to graph EKM (gEKM), which simultaneously utilizes both feature vectors and pairwise relations between data points. Specifically, gEKM integrates the proposed Elastic K-means with Normalized Cut clustering. In order to evaluate the effectiveness of proposed EKM and gEKM, we implement them on six benchmark datasets. The promising experimental results demonstrate the benefit of the proposed Elastic K-means models.

Elastic K-means and related works

K-means

Given data matrix $X = (x_1, \dots, x_n) \in \mathbb{R}^{p \times n}$ denoting n data points. The objective of K-means clustering is to find the cluster centroids $c_k, k = 1, \dots, K$ by minimizing the following cost function:

$$J_{km} = \sum_{i=1}^n \sum_{k=1}^K \|x_i - c_k\|^2 \tag{1}$$

Our work starts with the observation that the above K-means clustering objective can be reformulated as below:

Proposition 1. *The objective function Eq (1) can be rewritten as:*

$$\begin{aligned} \min_G \quad & J = \|X - XGG^T\|^2 \\ \text{s.t.} \quad & G^T G = \mathcal{N}^{-1}, \quad G_{ik} \in \{0, 1/\sqrt{n_k}\}. \end{aligned} \tag{2}$$

$\mathcal{N} = \text{diag}(n_1, \dots, n_k)$ where $n_i, (i = 1, \dots, k)$ denotes the number of data points in each cluster. Note that $\|A\|$ denotes the Frobenius norm of a matrix A . And $G \in \mathbb{R}^{n \times K}$ is the *normalized* cluster indicators, i.e.,

$$G_{i,k} = \begin{cases} 1/\sqrt{n_k} & \text{if } x_i \in c_k, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Proof of Proposition 1.

Let us introduce the standard cluster indicator

$$\tilde{G}_{i,k} = \begin{cases} 1 & \text{if } x_i \in c_k, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Then, Eq (1) can be equivalently written as:

$$\begin{aligned} \min_{C, \tilde{G}} \quad & J_{km} = \|X - C\tilde{G}^T\|^2 \\ \text{s.t.} \quad & \tilde{G}^T \tilde{G} = I, \quad \tilde{G}_{ik} \in \{0, 1\} \end{aligned} \tag{5}$$

where $C = (c_1, c_2, \dots, c_K) \in \mathbb{R}^{p \times K}$ are cluster centroids. The orthogonality of G is preserved in a relaxation of class indicator matrix [21]. Now, the cluster centroids can be written as $c_k = X\tilde{g}_k/n_k$, where \tilde{g}_k is the k -th column of \tilde{G} , or equivalently, $C = (c_1, \dots, c_K) = X\tilde{G}\mathcal{N}^{-1}$.

Thus:

$$\|X - C\tilde{G}^T\|^2 = \|X - X\tilde{G}\mathcal{N}^{-1}\tilde{G}^T\|^2 = \|X - XGG^T\|^2 \tag{6}$$

where $G = \tilde{G}\mathcal{N}^{-1/2}$ absorbs the unknown \mathcal{N} .

Elastic K-means

Proposition 1 provides an natural way to form the Elastic K-means. First, the orthogonal constraint $G^T G = I$ ensures that one data point belongs to only one cluster as in [21, 30]. We **relax** this constraint, so that each data point can belong to multiple clusters. Because the formulation Eq (2) is a matrix decomposition form, the constraint $G^T G = I$ will be satisfied approximately.

Second, since the mixed signs of G may deviate the results severely from the true solution of clustering [21, 30], we relax the discrete constraint $G_{ik} \in \{0, 1/\sqrt{n_k}\}$, to a simple nonnegativity constraint $G \geq 0$. Thus, the final model is formulated as:

$$\begin{aligned} \min_G \quad & J_{ekm} = \|X - XGG^T\|^2 \\ \text{s.t.} \quad & G \geq 0 \end{aligned} \tag{7}$$

Note that the input data X has mixed signs.

We call this Elastic K-means (EKM). (1) This model is invariant w.r.t. $X \rightarrow \beta X, \beta \in \mathfrak{R}$ is any constant. This is the same as the standard K-means. (2) Similar to K-means, the expected clustering centroids are adopted during the updating iterations. (3) different from the traditional K-means, the elastic indicator (interpreted as the posterior probability which is a fraction) performs a smooth convergency during updating. In fact, all the NMF-based methods we discussed above have K-means clustering interpretations when the factor G is orthogonal ($G^T G = I$), which in turn means they can be regarded as the relaxations of K-means.

Related works

Our goal is to study the case where a data point is not restricted to belong to a single cluster. Three related clustering models are (A) Gaussian mixtures [31, 32], (B) Fuzzy C-means [33–35], (C) Fuzzy K-means [12, 14, 36, 37]. In general, Gaussian mixtures works only for low-dimensional data, typically $d \leq 10$, therefore they are not suitable for high dimensional data. Fuzzy C-means solves

$$\min_{u, c_k} J_{fcm} = \sum_{i=1}^n \sum_{k=1}^K u_{ik}^m \|x_i - c_k\|^2 \tag{8}$$

where u_{ik} is the membership of data point x_i belonging to cluster k and $\sum_k u_{ik} = 1$.

Our model differs from Fuzzy C-means on two aspects: (1) Cluster distribution of single data point x_i is implicit in our model, whereas it is explicit in Fuzzy C-means. (2) The requirement that the posterior probability $\sum_k u_{ik} = 1$ for any data point x_i , but the actual distribution used in Eq (8) is $u_{ik}^m, m = 2$ in most cases. Thus the normalization of the actual cluster distribution is $z_i = \sum_k u_{ik}^m < 1$ and z_i is different for different data point x_i . This theoretical/conceptual drawback for Fuzzy C-means is not present in our Elastic K-means. Fuzzy K-means solves

$$\min_{w, c_k} J_{fkm} = \sum_{i=1}^n \sum_{k=1}^K w_{ik} \|x_i - c_k\|^2 \tag{9}$$

where w_{ik} is the membership of data point x_i belonging to cluster k and $\sum_k w_{ik} = 1$. The main difference between Fuzzy K-means and our method is, the soft/fuzzy capability is achieved by

exploring the corresponding weight w_{ik} (or u_{ik} in Fuzzy C-means) for each data point x_i , while our Elastic K-means achieves this via the soft/fuzzy cluster indicator G , which is more explicit. Note that, the concept of soft/fuzzy K-means (or C-means) has been mentioned in a group of literatures [38–42], which also have the fuzzy/soft capability for data clustering. However, these soft clustering methods are generally derived from Fuzzy C-means or Fuzzy K-means and their variants, which are essentially different from the proposed Elastic K-means model.

Because in our model, G is nonnegative. In some sense, our model also relates to NMF. There exists a very broad category of work along NMF direction. We refer to a recent survey [43].

Computational algorithm and analysis

Algorithm

An effective updating algorithm can be derived to solve Elastic K-means problem. The algorithm iteratively updates the current solution as follows:

$$G_{ik} \leftarrow G_{ik} \sqrt[4]{\frac{(2AG + BGG^T G + GG^T BG)_{ik}}{(2BG + AGG^T G + GG^T AG)_{ik}}} \tag{10}$$

The pseudo codes of EKM is illustrated as **Algorithm 1**.

Algorithm 1: Elastic K-means Algorithm

Input: data X

- 1 (1) Initialize G^0 .
- 2 Construct the indicator G : $G_{ik} = 1$ if x_i belongs to cluster k . otherwise, $G_{ik} = 0$
- 3 (2) Update G
- 4 **while not converged**
- 5 $G_{ik} \leftarrow G_{ik} \cdot \sqrt[4]{\frac{(2AG+BGG^T G+GG^T BG)_{ik}}{(2BG+AGG^T G+GG^T AG)_{ik}}}$
- 6 where
- 7 $A = (|X^T X|_{ik} + (X^T X)_{ik}) / 2$
- 8 $B = (|X^T X|_{ik} - (X^T X)_{ik}) / 2$
- 9 end

The convergence of the proposed algorithm can be found in the supplementary.

Updating algorithms for quartic models

There are many matrix models involving matrix variables to the 4th power. The simplest is:

$$\min_G \|W - GG^T\|_F^2 \tag{11}$$

where $W = W^T$ is symmetric. A number of papers claim that the updating algorithm for this model is

$$G_{ik} \leftarrow G_{ik} \frac{(WG)_{ik}}{(GG^T G)_{ik}} \tag{12}$$

In fact, this updating algorithm does not guarantee the monotonic decrease of the objective function value. Using the inequality of quartic matrix of Eq (S15) in the supplementary, we

can easily prove that the following updating algorithm;

$$G_{ik} \leftarrow G_{ik} \sqrt[4]{\frac{(WG)_{ik}}{(GG^T G)_{ik}}} \tag{13}$$

which guarantees decrease of objective, and thus the convergence.

Benefit of Elastic K-means

One main drawback of standard K-means is that each data point is assigned to a single cluster (hard clustering), but in real data, many data points are somewhere in-between different clusters centers (such as points 1, 2, 3 shown in Fig 1). One clear benefit of the proposed Elastic K-means is that these ambiguous data points are assigned into several nearby clusters, i.e., their posterior probabilities of cluster assignment are nearly evenly distributed.

Fig 1 shows the EKM results on one 2D dataset of 3 randomly generated Gaussian clusters. The EKM cluster results are indicated as red, green and blue stars respectively. Point 4, 5, 6

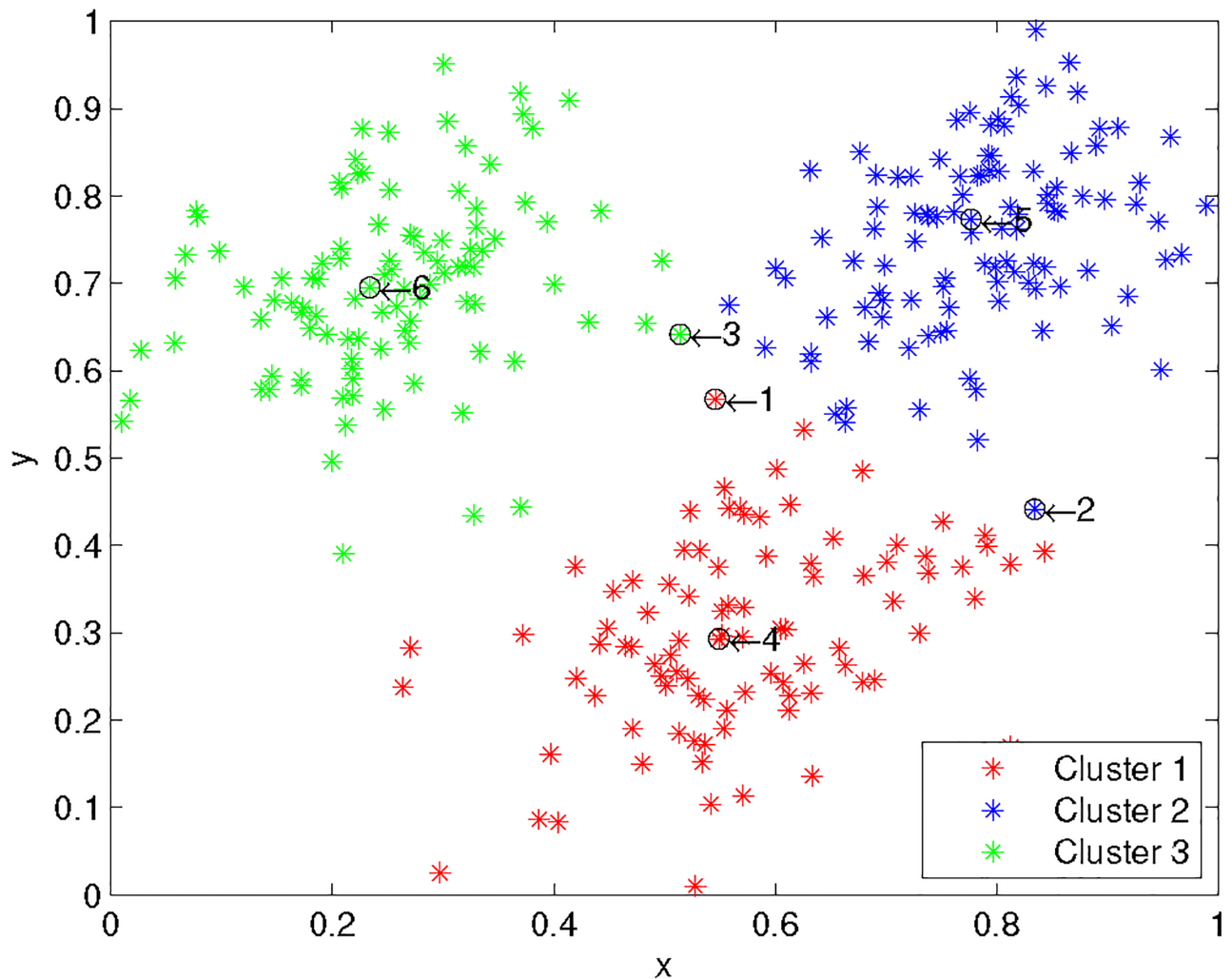


Fig 1. EKM clustering results on 2 dimensional 3 Gaussian clusters. x and y axes denote the first and the second dimension respectively.

<https://doi.org/10.1371/journal.pone.0188252.g001>

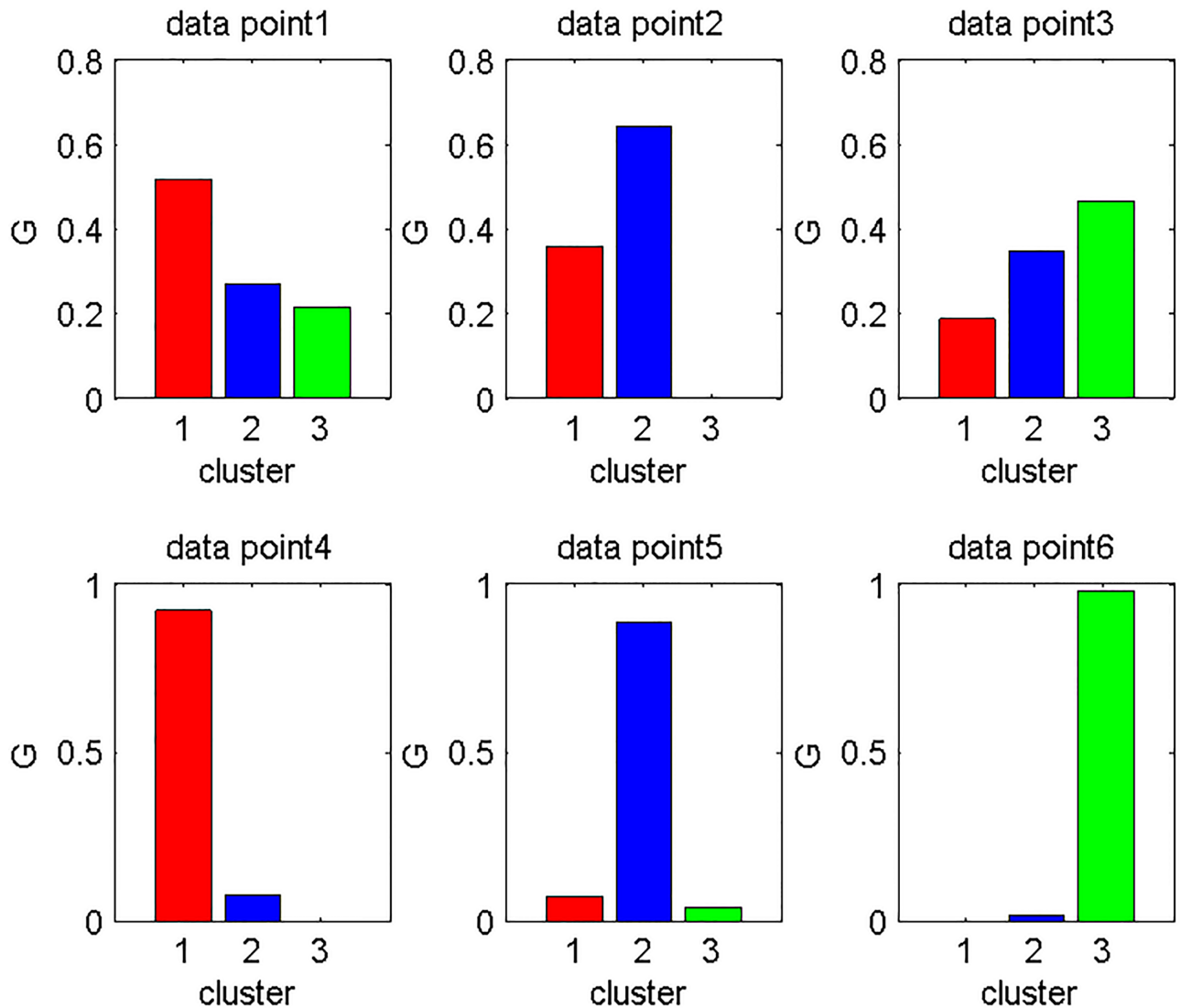


Fig 2. An example of the posterior G of 3 ambiguous data points and 3 sharp data points on above EKM clustering results.

<https://doi.org/10.1371/journal.pone.0188252.g002>

clearly belong to their assigned clusters. This fact is correctly encoded in the EKM model, as shown in the posterior probabilities in Fig 2: they are sharply concentrated on a single cluster.

On the other hand, points 1, 2, 3 are ambiguous: they are in-between different cluster centers. A good ‘elastic’ clustering model should assign these points to nearby clusters with nearly-equal posterior probabilities. Indeed, this desirable fact is correctly encoded in the EKM model, as shown in the posterior probabilities in Fig 2: they are nearly evenly distributed on nearby clusters.

Posterior Probability. In this and following sections, we normalize G to

$$G_{ik} \leftarrow G_{ik} / \sum_k G_{ik} \tag{14}$$

so that $\sum_k G_{ik} = 1$ for data point x_i .

A natural question is: how to identify those ambiguous data points such points 1, 2, 3 in Fig 2. For this purpose, we need to see how the posterior probability

$$p_i = G_i = (G_{i1}, G_{i2}, \dots, G_{ik}) \tag{15}$$

for data point x_i is distributed.

We use a useful quality to indicate approximately how sharply the distribution is peaked around the highest cluster. This is the gap in the posterior distribution: the difference between the highest and second highest peaks:

$$\Delta_i = (G_{ik_1} - G_{ik_2}) / G_{ik_1} \tag{16}$$

where

$$k_1 = \arg \max_{1 \leq k \leq K} G_{ik}, k_2 = \arg \max_{1 \leq k \leq K, j \neq k_1} G_{ik} \tag{17}$$

The idea is simple. (A) If a data point is well inside a cluster, such as points 4, 5, 6 in Fig 2, the highest peak is large and the second highest peak is small. Thus the gap is big. (B) If a data point is in-between different clusters, such as points 1, 2, 3, the difference between the highest peak and the second highest peak is small. Thus the gap is small. From these situations, the gap is useful indication of how sharply the posterior probability is distributed.

Fig 3 shows the distribution of the gap Δ_i of the 300 data points shown in Fig 1. Noted that, the lager Δ_i , the sharper that the data point belongs to a single class. Obviously, there are many ambiguous data points which can be easily detected using the gap. These studies indicate that Elastic K-means is able to detect the fuzzy/soft characteristics of data clustering.

Integrating Elastic K-means and normalized cut

EKM uses only vector (sometimes called attribute) data X . In many applications, the input data consists of both vector data X and similarity data W (also called graph data since W_{ij} represents the similarity between data points i and j). For this situation, we can easily extend EKM to utilize the similarity data W for more effective elastic clustering.

We incorporate the similarity data through the Normalized Cut [18, 19] formalism. This extension form a single clustering formulation for input data with both vector and similarity data.

The Ncut (Normalized Cut) is defined as:

$$\begin{aligned} \max_Q \quad & Tr(Q^T D^{-\frac{1}{2}} W D^{-\frac{1}{2}} Q) \\ \text{s.t.} \quad & Q^T Q = I, Q \in \{0, 1\} \end{aligned} \tag{18}$$

where $D = \text{diag}(d_1, \dots, d_n)$, $d_i = \sum_j W_{ij}$.

Proposition 2. Optimization problem Eq (18) is equivalent to

$$\begin{aligned} \min_Q \quad & \|S - QQ^T\|^2 \\ \text{s.t.} \quad & Q^T Q = I, Q \in \{0, 1\}. \end{aligned} \tag{19}$$

with the same constraints of Eq (18), where $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

Proof. Because $\|S - QQ^T\|^2 = Tr(S^2 - 2Q^T S Q + QQ^T Q Q^T)$, the first term and the last term are both constant due to $Q^T Q = I$. The minimization of $-2Q^T S Q$ is the same as Eq (18).

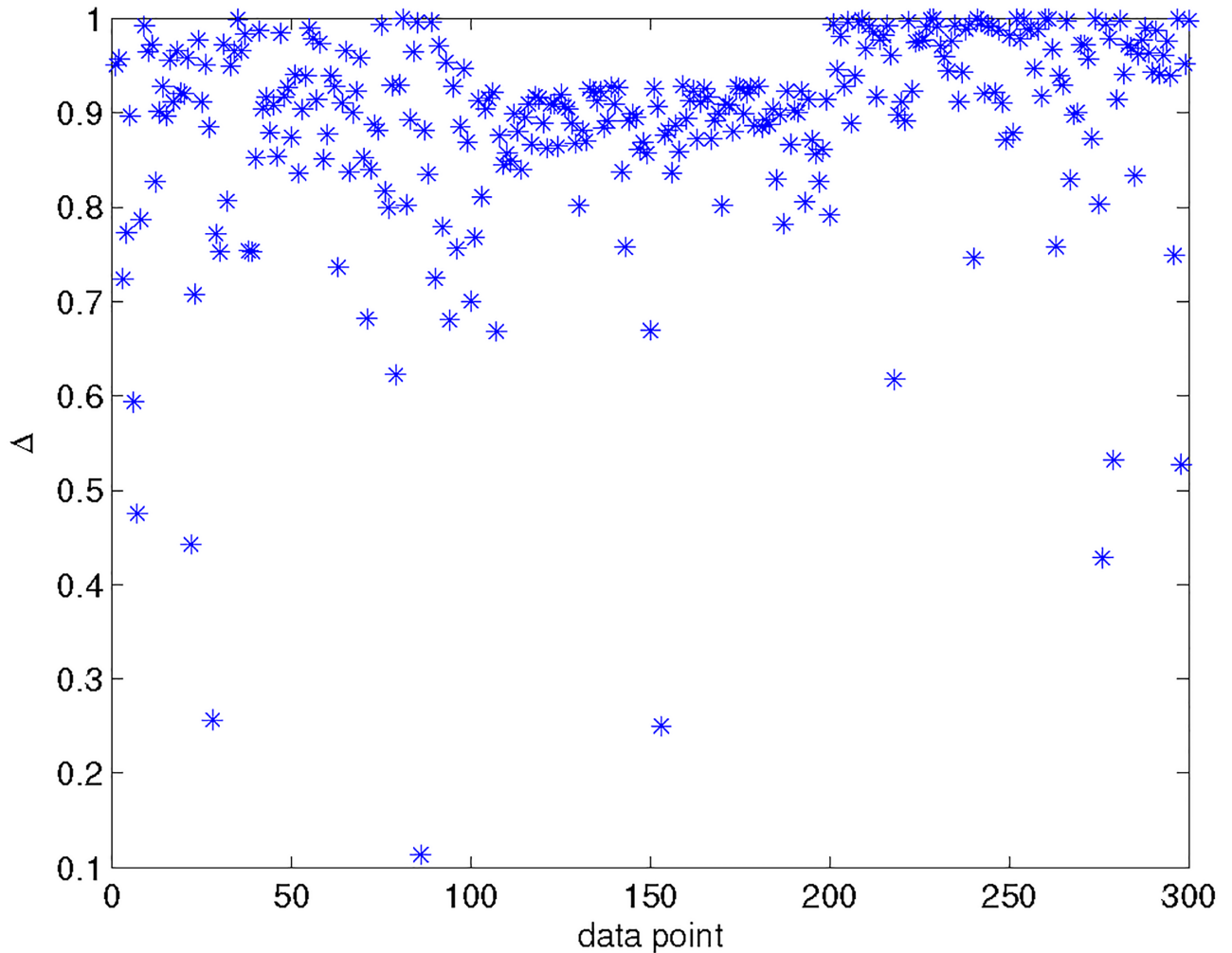


Fig 3. The gap in the posterior distribution: The difference between the highest and second highest peaks.

<https://doi.org/10.1371/journal.pone.0188252.g003>

In the same manner as EKM in Eq (7), we relax the orthogonal constraint $Q^T Q = I$ to be nonnegative, then Eq (19) can be rewritten as:

$$\begin{aligned} \min_Q \quad & \|S - QQ^T\|^2 \\ \text{s.t.} \quad & Q \geq 0 \end{aligned} \tag{20}$$

Since Q is also the cluster indicators as G in EKM, we can integrate EKM and Ncut by simplifying/approximating these constraints into $G \geq 0$. The above model is now expressed in a simplified way:

$$\begin{aligned} \min_G \quad & 2\|X - XGG^T\|^2 + \alpha \|S - GG^T\|^2 \\ \text{s.t.} \quad & G \geq 0 \end{aligned} \tag{21}$$

We call this model as graph EKM (simplified as gEKM in the following parts) since Ncut is a popular approach for graph data clustering. It naturally incorporates both the feature/attribute data X and pairwise relations W .

Analysis

Because the model is invariant with $X \rightarrow \beta X$, let $\beta = 1/\|X\|$, which implies the input X is normalized: $\|X\|^2 = \sum_{ij} X_{ij}^2 = 1$.

The above objective can be written as:

$$J = Tr[2X^T X + \alpha(S^2 - I) - 2G^T(X^T X + \alpha S)G] \tag{22}$$

The first 2 terms are independent of G . Thus the above optimization becomes

$$\begin{aligned} \min_G Tr G^T(X^T X + \alpha S)G \\ s.t. G \geq 0 \end{aligned} \tag{23}$$

where X is normalized.

This expression reveals an important **insight** that the gEKM model is a discrete combinatorial optimization.

Algorithm 2

The optimization problem of Eq (21) can be solved by the following iterative algorithm:

$$G_{ik} \leftarrow G_{ik} \sqrt[4]{\frac{(2AG + BGG^T G + GG^T BG + 2\alpha_1 SG)_{ik}}{(2BG + AGG^T G + GG^T AG + 2\alpha_1 GG^T G)_{ik}}} \tag{24}$$

where $\alpha_1 = \alpha * \|X\|^2$, and A, B are defined in **Algorithm 1**. The proof of convergence can be similarly established as **Algorithm 1**.

Experiments

We perform experiments on following six benchmark datasets described as **Table 1**.

AT & T (<http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>) face dataset contains ten different images of each 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expression and facial details (glass/no glass). All images were taken against a dark homogeneous background with the subjects in an upright, frontal, position. We stretch each image into a vector.

USPS (<http://www.cs.nyu.edu/roweis/data.html>) is a handwritten digit (1–10) database and we select 1000 images (100 images for every digit).

MNIST (<http://www.cs.nyu.edu/roweis/data.html>) is a handwritten digit database. Each image is centered (according to the center of mass of the pixel intensities) on a 28x28 grid. In our experiments, we randomly choose 1000 images (i.e., each digit has 100 images). We reshape each image into one vector.

Table 1. Dataset names, number of data samples, data dimension and number of classes of the 6 benchmark datasets.

Dataset	#Sample	#Dimension	#Class
AT&T	400	1024	40
USPS	9298	256	10
MNIST	1000	784	10
COIL20	1440	1024	20
Isolet1	1560	617	26
BinAlph	1014	320	26

<https://doi.org/10.1371/journal.pone.0188252.t001>

COIL20 (<http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php>) contains 20 objects. Each image of the same object is taken 5 degrees apart as the object is rotated on a turntable and each object has 72 images. The size of each image is 3232 pixels, with 256 gray levels per pixel. Each image is represented by a 1024 dimensional vector.

Isolet1 (<http://archive.ics.uci.edu/ml/datasets/ISOLET>) dataset was generated by 150 subjects spoken the name of each letter of the alphabet twice. The speakers are grouped into sets of 30 speakers each, and are referred to as isolet1, isolet2, isolet3, isolet4, isolet5. The features include spectral coefficients, contour features, sonorant features, pre-sonorant features, and post-sonorant features. In our experiment, we utilize subset isolet1 only.

BinAlpha (<http://www.cs.nyu.edu/roweis/data.html>) contains 26 binary hand-written alphabets and we select 30 images for every alphabet. We stretch each image into one vector.

Clustering accuracy is used to measure the performances. Once the clustering solution is computed, the confusion matrix is calculated using the Hungarian algorithm [44] is employed to match obtained clusters with the ground truth classes. The accuracy of this matching is computed as the clustering accuracy.

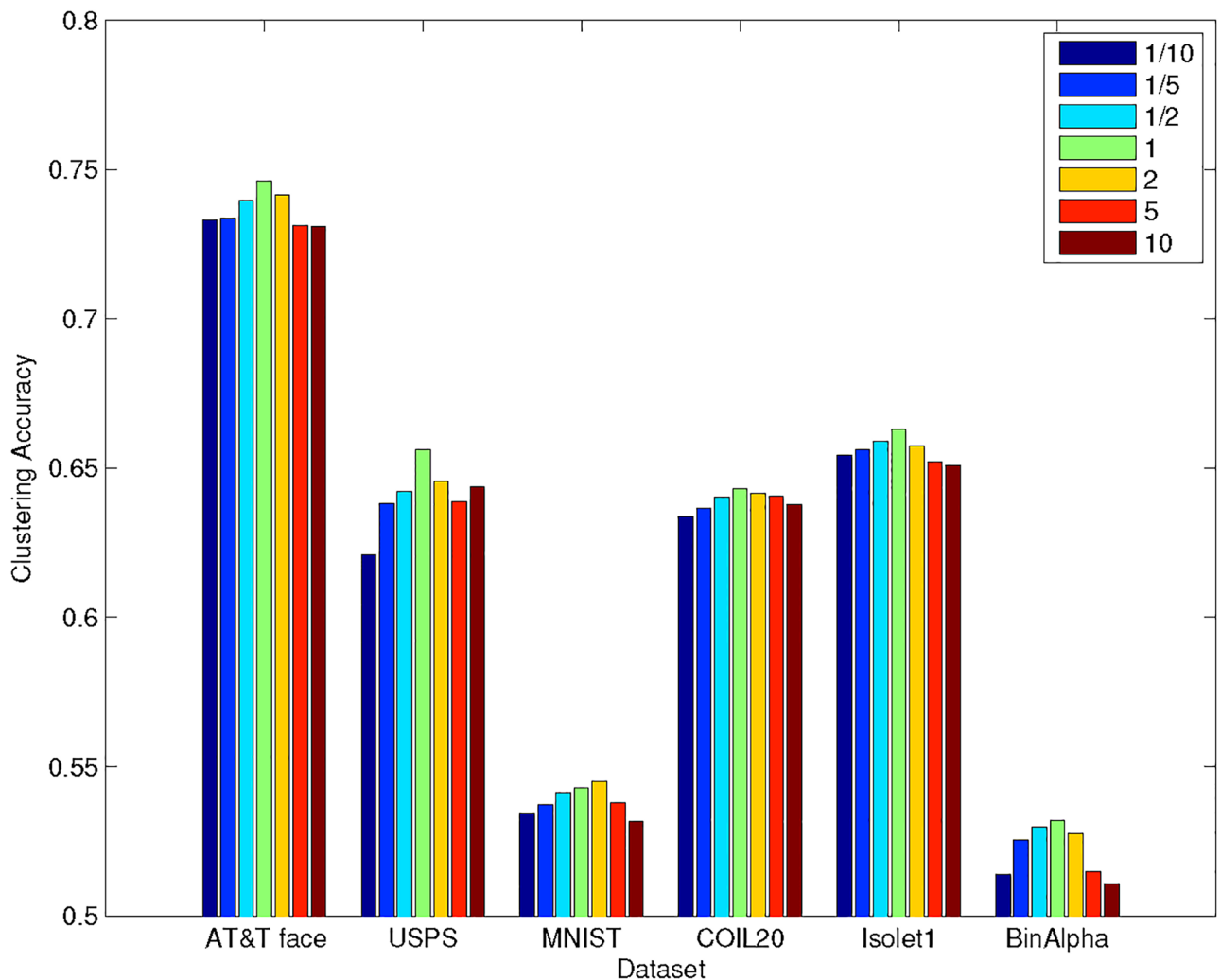


Fig 4. Clustering accuracy for gEKM against α on six datasets in full dimension.

<https://doi.org/10.1371/journal.pone.0188252.g004>

In all experiments, for each dataset, we first run 20 K-means clustering with random starts. We pick the best (lowest clustering objective function value) result of these 20 runs, record the corresponding clustering accuracy as the result of K-means.

Starting from this K-means result, we run NMF [11], FCM [34], MinMaxCut [21], EKM, gEKM, and record the corresponding results. For Ncut [18], we run K-means in the eigenspace 20 times with random starts, and pick the best (lowest K-means clustering objective function value) result of these 20 runs.

Choice of parameter α

The clustering accuracy of gEKM depends on parameter α . If $\alpha > 1$, the similarity part (2nd term of Eq (21)) is more important. If $\alpha < 1$, the vector attribute part (first term of Eq (21)) is more important. For this reason, the proper range of choice for α is [1/10, 1/5, 1/2, 1, 2, 5, 10]. We show the clustering results against α on all six datasets in Fig 4 (here X is in full dimension).

We observe that (1) The results are not very sensitive to α . (2) $\alpha = 1$ generally gives best results, which in turn means the vector attribute and the similarity part have more or less the balanced contribution. From these observations, we fix α to 1 in the following experiments.

Convergency visualization

Fig 5 visually illustrates the changes of the objective functions of Eqs (7) and (21), which demonstrates the convergency of the objective functions.

Results on clustering

We compare our methods, EKM and gEKM to K-means, NMF [11], FCM [34], Ncut [18] and MixMaxCut [21] on the six datasets. For EKM, Ncut [18] and MinMaxCut [21], we compute

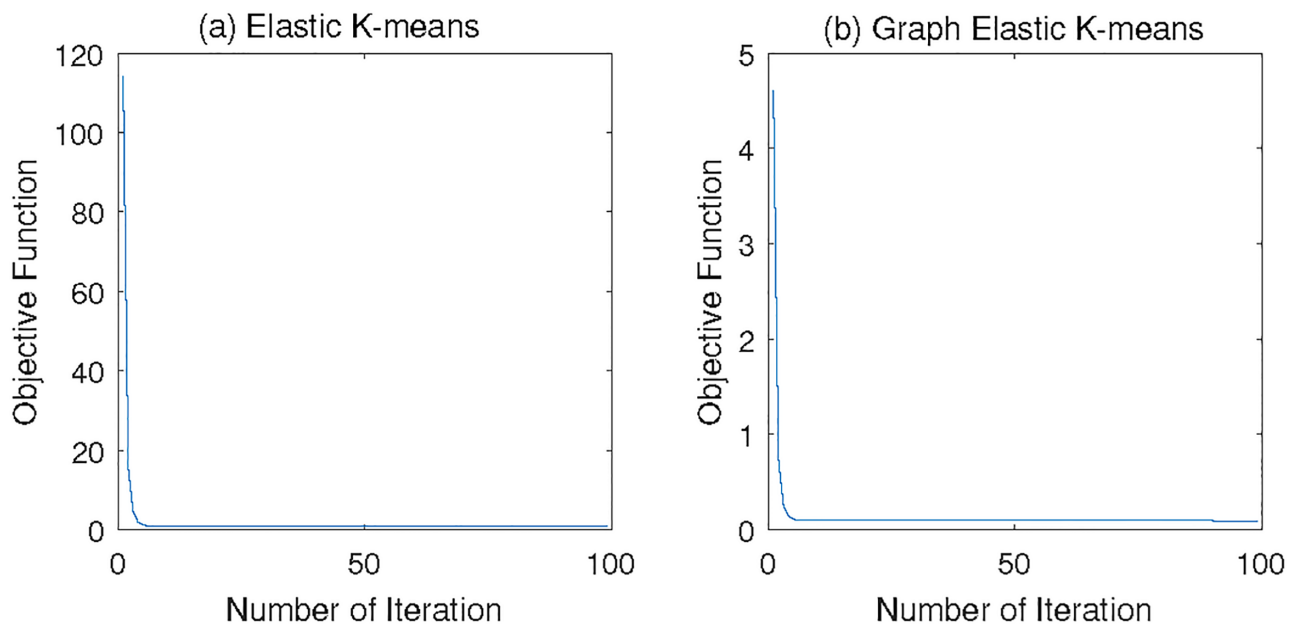


Fig 5. Changes of the objective function of the EKM and gEKM in 100 iterations.

<https://doi.org/10.1371/journal.pone.0188252.g005>

Table 2. Clustering accuracy of six datasets in full dimension.

Dataset	Kmeans	FCM	NMF	Ncut	MinMaxCut	EKM	gEKM
AT&T	0.715	0.701	0.721	0.731	0.730	0.733	0.746
USPS	0.618	0.626	0.624	0.644	0.646	0.621	0.656
MNIST	0.494	0.501	0.518	0.531	0.532	0.534	0.545
COIL20	0.625	0.615	0.631	0.638	0.635	0.634	0.643
Isolet1	0.626	0.624	0.628	0.651	0.645	0.654	0.663
BinAlph	0.495	0.511	0.502	0.511	0.514	0.512	0.532

<https://doi.org/10.1371/journal.pone.0188252.t002>

Table 3. Clustering accuracy of six datasets in PCA subspace when p = 50.

Dataset	Kmeans	FCM	NMF	Ncut	MinMaxCut	EKM	gEKM
AT&T	0.713	0.713	0.715	0.736	0.731	0.734	0.748
USPS	0.604	0.619	0.621	0.627	0.632	0.636	0.637
MNIST	0.521	0.523	0.524	0.545	0.537	0.536	0.551
COIL20	0.620	0.631	0.627	0.633	0.640	0.652	0.658
Isolet1	0.626	0.627	0.626	0.641	0.635	0.646	0.653
BinAlph	0.498	0.510	0.518	0.534	0.533	0.541	0.553

<https://doi.org/10.1371/journal.pone.0188252.t003>

Table 4. Clustering accuracy of six datasets in PCA subspace when p = 100.

Dataset	Kmeans	FCM	NMF	Ncut	MinMaxCut	EKM	gEKM
AT&T	0.711	0.701	0.711	0.737	0.731	0.738	0.749
USPS	0.608	0.625	0.625	0.632	0.622	0.635	0.641
MNIST	0.516	0.521	0.525	0.534	0.535	0.539	0.546
COIL20	0.618	0.625	0.630	0.633	0.634	0.637	0.650
Isolet1	0.628	0.623	0.627	0.637	0.639	0.645	0.656
BinAlph	0.481	0.515	0.518	0.523	0.522	0.528	0.534

<https://doi.org/10.1371/journal.pone.0188252.t004>

the graph similarity W as:

$$W_{ij} = \exp(-\|x_i - x_j\|^2 / cd^2) \tag{25}$$

where $c = 0.7$ and d is the average distances between each x_i and its 7 nearest neighbors. We set $W_{ii} = 0$ and construct $S = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.

Table 2 reports the clustering accuracy on the six datasets using K-means, NMF, FCM, Ncut, MinMaxCut, EKM and gEKM. One can see that (1) Ncut and MinMaxCut consistently

Table 5. Clustering accuracy of six datasets in PCA subspace when p = 150.

Dataset	Kmeans	FCM	NMF	Ncut	MinMaxCut	EKM	gEKM
AT&T	0.717	0.702	0.702	0.735	0.720	0.722	0.741
USPS	0.605	0.612	0.627	0.631	0.623	0.637	0.638
MNIST	0.519	0.511	0.528	0.537	0.539	0.541	0.551
COIL20	0.607	0.622	0.629	0.631	0.633	0.638	0.644
Isolet1	0.627	0.622	0.631	0.642	0.640	0.645	0.658
BinAlph	0.485	0.512	0.519	0.526	0.531	0.536	0.538

<https://doi.org/10.1371/journal.pone.0188252.t005>

Table 6. Clustering accuracy of six datasets in PCA subspace when p = 200.

Dataset	Kmeans	FCM	NMF	Ncut	MinMaxCut	EKM	gEKM
AT&T	0.712	0.708	0.702	0.731	0.725	0.723	0.736
USPS	0.610	0.616	0.625	0.629	0.631	0.641	0.643
MNIST	0.512	0.503	0.518	0.525	0.532	0.539	0.545
COIL20	0.623	0.611	0.621	0.626	0.630	0.636	0.643
Isolet1	0.619	0.624	0.625	0.631	0.629	0.633	0.644
BinAlpha	0.477	0.512	0.515	0.522	0.525	0.528	0.537

<https://doi.org/10.1371/journal.pone.0188252.t006>

outperform the existing NMF, FCM and Kmeans methods. (2) The proposed EKM beats Ncut and MinMaxCut in most of the cases. (3) MinMaxCut and Ncut perform nearly to proposed EKM and slightly better in some cases. (4) The integrated gEKM consistently outperforms the other six methods including MinMaxCut and Ncut which verifies the effectiveness of the

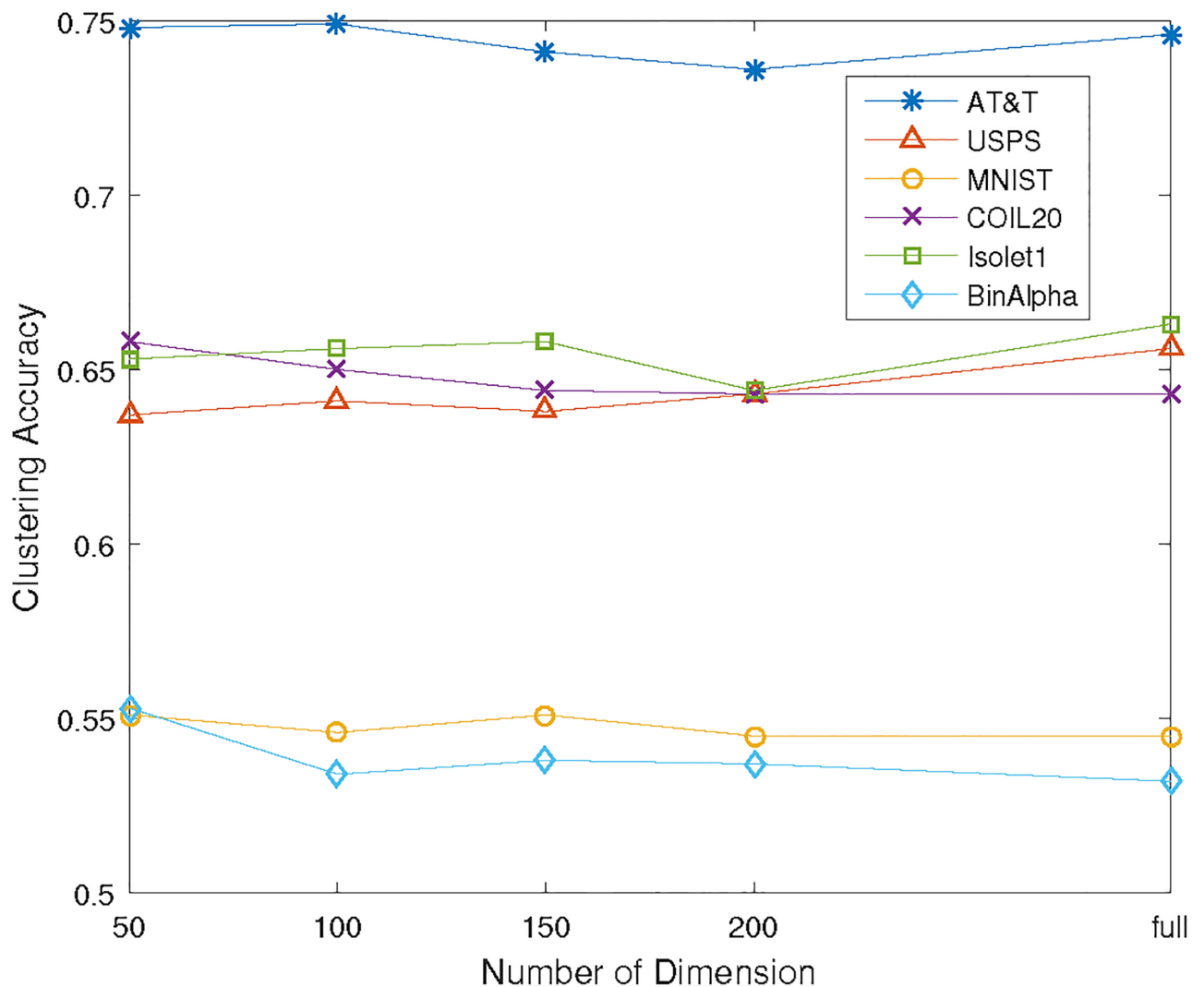


Fig 6. Clustering accuracy of gEKM against the number of dimension of PCA subspace on six datasets. where “full” on the x-axis denotes the full dimension of the datasets as indicated on Table 1.

<https://doi.org/10.1371/journal.pone.0188252.g006>

integration. In order to elaborate the performance of proposed methods, we evaluate the clustering experiments in different PCA subspace with the dimension of $p = 50, 100, 150, 200$ on the six datasets. The obtained clustering accuracies are given in Tables 3 to 6. First of all, the results on different PCA subspaces are consistent as the full space as we discussed above. Furthermore, as shown in Fig 6, the clustering results on PCA subspace can achieve competitive or even slight better than the full space, which implies the effectiveness of subspace discovery.

Conclusion

In this paper, we firstly propose a Elastic K-means(EKM) framework which provides a elastic clustering solution and prove that the correctness and the convergency of the updating algorithm. Secondly, an integrated framework specified by the combination of EKM and Normalized Cut (gEKM) is proposed to take into account of both attribute data and pairwise relations and the choice of combination parameter is particularly analysed. The experimental results on six benchmark datasets demonstrate the proposed EKM leads to the better clustering accuracy than K-means as well as previous NMF-based algorithms and the popular Fuzzy C-means and Fuzzy K-means. The gEKM has been also been tested to have satisfied performance on data clustering.

Supporting information

S1 File. Supplementary materials.
(PDF)

S2 File. Supplementary materials.tex file.
(TEX)

Author Contributions

Conceptualization: Chris Ding.

Formal analysis: Xuehan Zhang.

Methodology: Bo Jiang.

Supervision: Aihua Zheng.

Validation: Yan Li.

Visualization: Xuehan Zhang.

Writing – original draft: Yan Li.

Writing – review & editing: Aihua Zheng, Xuehan Zhang, Chris Ding.

References

1. Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, et al. Orthology detection combining clustering and synteny for very large datasets. *PLoS One*. 2014; 9(8):e105015. <https://doi.org/10.1371/journal.pone.0105015> PMID: 25137074
2. Huang J, Nie F, Huang H, Ding C. Robust Manifold Nonnegative Matrix Factorization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2014; 8(3):11:1–11:21. <https://doi.org/10.1145/2601434>
3. Melin P, Castillo O. A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition. *Applied Soft Computing*. 2014; 21:568–577. <https://doi.org/10.1016/j.asoc.2014.04.017>

4. Yang H, Seoighe C. Impact of the Choice of Normalization Method on Molecular Cancer Class Discovery Using Nonnegative Matrix Factorization. *PLoS one*. 2016; 11(10):e0164880. <https://doi.org/10.1371/journal.pone.0164880> PMID: 27741311
5. Ji Z, Liu J, Cao G, Sun Q, Chen Qg. Robust spatially constrained fuzzy c-means algorithm for brain MR image segmentation. *Pattern Recognition*. 2014; 47:2454–2466. <https://doi.org/10.1016/j.patcog.2014.01.017>
6. Wu S, Feng X, Zhou W. Spectral clustering of high-dimensional data exploiting sparse representation vectors. *Neurocomputing*. 2014; 135:229–239. <https://doi.org/10.1016/j.neucom.2013.12.027>
7. Lakizadeh A, Jalili S. BiCAMWI: A Genetic-Based Biclustering Algorithm for Detecting Dynamic Protein Complexes. *PLoS one*. 2016; 11(7):e0159923. <https://doi.org/10.1371/journal.pone.0159923> PMID: 27462706
8. Edgardo M, Daniel T, Javier S, Carlos G, Francisco T, Alberto P. NMF-mGPU: non-negative matrix factorization on multi-GPU systems. *BMC Bioinformatics*. 2015; 16:1–12.
9. Biswas AK, Gao JX, Zhang B, Wu X. NMF-Based LncRNA-Disease Association Inference and Bi-Clustering. In: *IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*; 2014. p. 97–104.
10. Zhang J, O'Reilly KM, Perry GL, Taylor GA, Dennis TE. Extending the functionality of behavioural change-point analysis with k-means clustering: a case study with the little penguin (*eudyptula minor*). *PLoS one*. 2015; 10(4):e0122811. <https://doi.org/10.1371/journal.pone.0122811> PMID: 25922935
11. Ding C, Li T, Jordan MI. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2010; 24:45–55. <https://doi.org/10.1109/TPAMI.2008.277>
12. Xu J, Han J, Xiong K, Nie F. Robust and Sparse Fuzzy K-Means Clustering. In: *IJCAI*; 2016. p. 2224–2230.
13. Liu H, Wu Z, Li X, Cai D, Huang TS. Constrained Nonnegative Matrix Factorization for Image Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012; 34:1299–1311. <https://doi.org/10.1109/TPAMI.2011.217>
14. Nie F, Zeng Z, Tsang IW, Xu D, Zhang C. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*. 2011; 22(11):1796–1808. <https://doi.org/10.1109/TNN.2011.2162000> PMID: 21965198
15. Qiao H. New SVD based initialization strategy for Non-negative Matrix Factorization. *Pattern Recognition Letters*. 2015; 63:71–77. <https://doi.org/10.1016/j.patrec.2015.05.019>
16. Hagen L, Kahng AB. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. 1992; 11:1074–1085. <https://doi.org/10.1109/43.159993>
17. Hu Z, Pan G, Wang Y, Wu Z. Spectral Sparse Representation for Clustering: Evolved from PCA, K-means, Laplacian Eigenmap, and Ratio Cut. *arXiv preprint*. 2014; arXiv:1403.6290.
18. Shi J, Malik JP. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000; 22:888–905. <https://doi.org/10.1109/34.868688>
19. Yan X, Guo J, Liu S, Cheng Xq, Wang Y. Clustering short text using Ncut-weighted non-negative matrix factorization. In: *CIKM'12 Proceedings of the 21st ACM international conference on Information and knowledge management*; 2012. p. 2259–2262.
20. Neumayer S, Efrat A, Modiano E. Geographic max-flow and min-cut under a circular disk failure model. *Computer Networks*. 2015; 77:117–127. <https://doi.org/10.1016/j.comnet.2014.10.026>
21. Nie F, Ding C, Luo D, Huang H. Improved minmax cut graph clustering with nonnegative relaxation. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer; 2010. p. 451–466.
22. Tenenbaum JB, Silva Vd, C LJ. A global geometric framework for nonlinear dimensionality reduction. *Science*. 2000; 290:2319–2323. <https://doi.org/10.1126/science.290.5500.2319> PMID: 11125149
23. Bowen GJ, West J, Miller C. IsoMAP: Isoscapes Modeling, Analysis and Prediction (version 1.0). The IsoMAP Project. 2012;.
24. Zhao Z, Chow TWS, Zhao M. M-Isomap: Orthogonal Constrained Marginal Isomap for Nonlinear Dimensionality Reduction. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society*. 2013; 43(1):180–191.
25. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*. 2000; 290(5500):2323–2326. <https://doi.org/10.1126/science.290.5500.2323> PMID: 11125150
26. Xiang S, Nie F, Pan C, Zhang C. Regression Reformulations of LLE and LTSA With Locally Linear Transformation. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of*

- the IEEE Systems Man & Cybernetics Society. 2011; 41:1250–1261. <https://doi.org/10.1109/TSMCB.2011.2123886>
27. Deng T, Deng Y, Shi Y, Zhou X. Research on Improved Locally Linear Embedding Algorithm. In: Bio-Inspired Computing—Theories and Application; 2014. p. 88–92.
 28. Zhang Zy, Zha Hy. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University (English Edition)*. 2004; 8:406–424. <https://doi.org/10.1007/s11741-004-0051-1>
 29. Wang J. Local tangent space alignment. In: *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer; 2011. p. 221–234.
 30. Yang Y, Shen HT, Nie F, Ji R, Zhou X. Nonnegative spectral clustering with discriminative regularization. In: *AAAI*; 2011. p. 2–4.
 31. Jordan M, Xu L. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*. 2008; 8(1):129–151.
 32. Polanski A, Marczyk M, Pietrowska M, Widlak P, Polanska J. Signal partitioning algorithm for highly efficient Gaussian mixture modeling in mass spectrometry. *PLoS one*. 2015; 10(7):e0134256. <https://doi.org/10.1371/journal.pone.0134256> PMID: 26230717
 33. Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*. 1984; 10(84):191–203. [https://doi.org/10.1016/0098-3004\(84\)90020-7](https://doi.org/10.1016/0098-3004(84)90020-7)
 34. Tang JR, Isa NAM, Chng ES. A Fuzzy-C-Means-Clustering Approach: Quantifying Chromatin Pattern of Non-Neoplastic Cervical Squamous Cells. *PLoS one*. 2015; 10(11):e0142830. <https://doi.org/10.1371/journal.pone.0142830> PMID: 26560331
 35. Bai C, Dhavale D, Sarkis J. Complex investment decisions using rough set and fuzzy c-means: an example of investment in green supply chains. *European journal of operational research*. 2016; 248(2):507–521. <https://doi.org/10.1016/j.ejor.2015.07.059>
 36. Bezdek JC. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*. 1980;(1):1–8. <https://doi.org/10.1109/TPAMI.1980.4766964> PMID: 22499617
 37. Li MJ, Ng MK, Cheung Ym, Huang JZ. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *IEEE transactions on knowledge and data engineering*. 2008; 20(11):1519–1534. <https://doi.org/10.1109/TKDE.2008.88>
 38. Yin X, Chen S, Hu E. Regularized soft K-means for discriminant analysis. *Neurocomputing*. 2013; 103(3):29–42. <https://doi.org/10.1016/j.neucom.2012.08.021>
 39. Bai X, Luo S, Zhao Y. Entropy based soft K-means clustering. *IEEE*; 2008.
 40. Kim J, Shim KH, Choi S. Soft Geodesic Kernel K-Means. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*; 2007. p. II-429–II-432.
 41. Yu YC, Wang JD, Zheng GS, Jiang Y. Distributed K-means based on Soft Constraints. *Journal of Software Engineering*. 2011;. <https://doi.org/10.3923/jse.2011.116.126>
 42. Yang T, Wang J. A Robust k-Means Type Algorithm for Soft Subspace Clustering and Its Application to Text Clustering. *Journal of Software*. 2014; 9(8). <https://doi.org/10.4304/jsw.9.8.2120-2124>
 43. Li T, Ding CH. Nonnegative Matrix Factorizations for Clustering: A Survey. In: *Data Clustering: Algorithms and Applications*; 2013. p. 149–176.
 44. Kuhn HW. The Hungarian Method for the Assignment Problem. In: *50 Years of Integer Programming 1958–2008*; 2010. p. 29–47.