MATH4983F Final Report
Human Face Classification with New Clustering Apporach

Chan Kin Hang

03/05/2019

1 Introduction

Clustering is the simplest topic in machine learning for designing an algorithm to classify the elements of a set into groups or clusters. The first algorithm is k-mean which is mentioned in 1957 [1]. Clustering algorithm is an unsupervised classification and hence the performance is much lower that that of supervised algorithm such as Support Vector Machine (SVM) and Neutral Network. However, it doesnt mean that it is useless. In bioinformatic fields, subtypes of cancer can be discovered by the use of clustering method which improve the diagnosis of patients with cancer. Nowadays, clustering is applied in industrial AI algorithm which act as an aide. For example, it is able to determine the best training set for a single neural network. Apart from that, in object recognition, k-mean based feature descriptor with SVM has excellent performance[2]. In this project, clustering method is used to classify human face in ORL (Olivetti Research Laboratory face database in Cambridge)[3]. The recall is 0.905 for known number of clusters, which is better than some papers. For the unknown number of clusters, it fails.

In this project, we first input the photo as data from ORL. Then, for each photo, the feature points are extracted by calculating the coordinates and feature vectors of each feature point through SIFT (Scale Invariant Feature Transform). Base on the pervious results, the matching number of each data can be found and presented as a dissimilarity matrix. Moreover, a new kernel function is applied on that matrix to form a modified dissimilarity matrix. Finally, it is clustered with PAM clustering with a known number of cluster k.

Moreover, for an ideal unsupervised clustering algorithm, finding the unknown number of cluster k is critical. This aim is achieved by applying the clustering validation which is proposed to evaluate the optimal solution of clustering validation measures. The details will be discussed after the clustering section.

2 Methodology

2.1 Clustering with known number of clusters k

2.1.1 Olivetti Research Laboratory face database

ORL which is one of the branch marks of face recognition involves 400 photos (92x112) of 40 people while 10 photos for each person. All photos are in dark homogenous background. The people took the photos in frontal position with various time which was lasting for 2 years in the whole research, and with different facial express and accessories each time. Besides, there are other branch mark databases such as Yale and Yale B and the characteristic of each of them are mentioned in Handbook of Face Recognition written by S .Li and A .Jain[4]. It involves no. of subjects, pose, illumination, facial expressions and time. The number of those characteristics which cannot be determined is highest in ORL and the number of subjects

(number of photos per person) is lowest. Hence, it may be the hardest database and is worth to adopted in this project.
As clustering is unsupervised, its power is limited. Hence, the input data is one fourth of ORL which contains all the photos of 10 people, so the total number of photos are 100. In order to analyze the whole ORL, the test will be conducted four times while the people involved will not be repeated in MATLAB. It will be processed for three more times. The final result is the average. Each photo is transformed into a 112 x 92 matrix as an input datum.

2.1.2 Calculating D by SIFT

Scale Invariant Feature Transform (SIFT) is an image descriptor for image-based matching and recognition developed by David Lowe (1999, 2004)[5]. In this paper, it is adopted as it can give a highly accurate precise feature extraction result base on the following reasons. The matching result is uncorrelated with image transformations. Also, the feature points of an image are extracted without any bias. Moreover, the coordinate of feature point is precise so that two points which are closed to each other can be distinguished. The feature vector of that point is accurate and precise. All the reasons can be explained by the formulas in SIFT.
In this project, image in the matrix form is the input. By using VLFeat (version 0.9.21), the interest/feature points are detected by the difference-of-Gaussians operator which involves the scale-normalized Laplacian:

$$DOG(x, y; s) \approx \frac{k^2 - 1}{2} \bigtriangledown^2_{norm} L(x, y; s) \tag{1}$$

Then the feature vector with 128 components of this point is computed by two formulas shown below:

$$arg \bigtriangledown L = atan2(Ly, Lx) \tag{2}$$

$$| \bigtriangledown L| = \sqrt{L_x^2 + L_y^2} \tag{3}$$

And its domain is 4 × 4 grid.
In matching process, the similarity of each image by each image is qualified by the matching number of image descriptors/feature vectors, which consider the minimum Euclidean distance between the vector and the other vector in other face image. There is the denoting, the vector $\vec{I_i} \in \mathbb{R}^{100}$ represents the matching numbers of the image i by each image, where each element I_{i_j} represents the matching number of image i with respect to image j. To avoid possibly ambiguous, the ratio the nearest distance to the second nearest distance is 0.8 in this program. Finally, a $n \times n$ dissimilarity matrix D can be created, where each entry $D_{ij} = I_{i_j}$ represents the matching number of image i and image j and each row i is equal to $\vec{I_i}$. n is the total image number.

2.1.3 Modified Dissimilarity Matrix D by new kernel function

In practice, D is not symmetric. However, it must be symmetric as the number of matching of face image descriptors between image i and j (D_{ij}) must be the same with image j and i (D_{ji}). In order to fix this problem and collect an accurate clustering result, it is modified by the following new kernel function.

$$D_{ij}{'} = k(I_i, I_j) = \frac{2Min(I_{i_j}, I_{j_i})}{I_{i_i} + I_{j_j}} \tag{4}$$

where I_{i_i} is the number of feature points in image i and I_{i_j} is the matching number of image i and image j. This matrix is called D which is symmetric and normalized.

The range of each entry $(D_{ij}{}')$ is between 0 and 1 where the value near to 0 implies that higher dissimilarity between the two images. Hence, the data which are similar will be near to each other in a $\mathbb{R}^{100}$ vector space. Then PAM can be applied on it effectively.
The reason of choosing the minimum M_{ij} between I_{i_j} or I_{j_i} is to ensure the possibility of this matching number M_{ij} which is larger than that of the real one is low while M_{ij} is relatively large. For the experimental way, the other options, maximum and average, are tested. The recalls of both results are lower than the minimum one. Figure 1 shows the modified dissimilarity matrices D for each data subset used in this project.

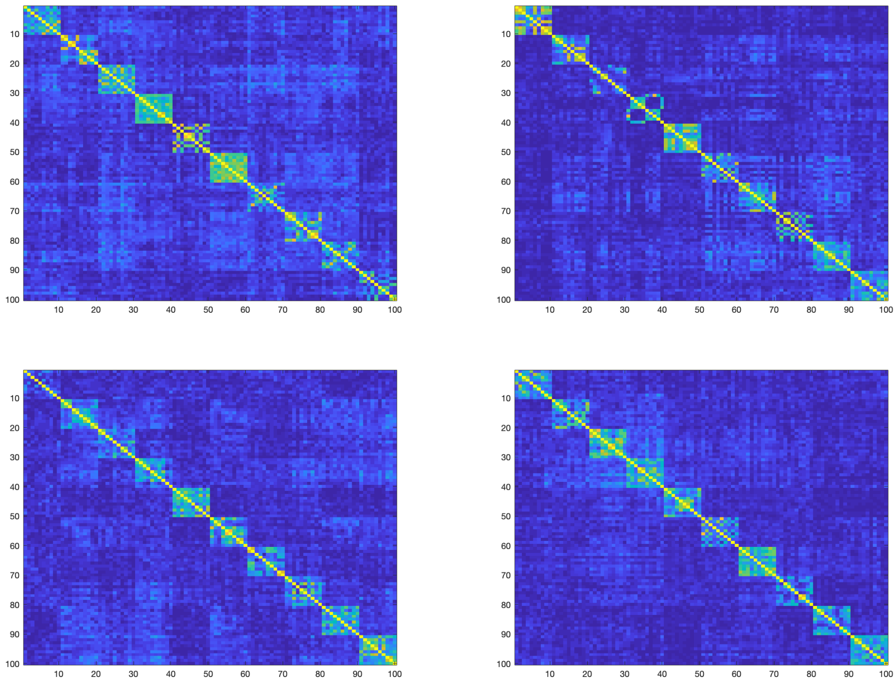

Figure 1: Modified dissimilarity matrices D' calculated from image features derived by SIFT and new kernel function for all images in each part of ORL face database

2.1.4 PAM

One of the clustering algorithms is PAM which is used to solve k-medoids problems. The idea behind is similar to k-means but PAM has better performance and can prevent outliers. However, it is less popular than k-mean.(It might because of the random order of labelling.) The details is shown below.

Initial phase:
1. Let U be the set such that the coordinates of all data points $u_1, u_2, ..., u_i, ..., u_{100}$ are elements belongs to U, where $u_i = \vec{I_i}$.
2. Given the number of cluster k.
3. Select k elements from U as medoids arbitrary $(i.e. u_a = m_1, u_b = m_2, ..., u_j = m_l, ..., u_p = m_k)$, which can prevent outliers.
4. Denote the distance function $d_{il} = d(u_i, m_l) = ||u_i - m_l||_2{}^2$, it is called square 2-norm, which is sensitive for comparing the differences between distances.
5. Calculate all the combination of d_{il}.
6. Denote set $M_1, M_2, ..., M_l, ..., M_k$ such that $m_1 \in M_1, m_2 \in M_2, ..., m_k \in M_k$.

For each point u_i, find the minimum distance $d_{i\,min}$ with one of the medoids m_l. Then $u_i \in m_l$ (i.e. For fixed i, $u_i \in m_l$, where $l = \underset{1 \le l \le k}{\arg\min}\, d_{il}$)
Swapping phase:
1. For M_n, select another point as medoids $m_n{}'$ arbitrary such that $m_n{}' \in M_n$ and $m_n{}' \neq m_n$. Repeat step 5 and denote the minimum distance $d_{i\,min}$ for each u_i is $d_{i\,min}{}'$.
2. Case 1: If $\sum_{i=1}^{n=100} d_{i\,min}{}' < \sum_{i=1}^{n=100} d_{i\,min}$, then m_n be the new medoids (called swapping) and repeat step 7 for M_{n+1}.
Case 2: If $\sum_{i=1}^{n=100} d_{i\,min}{}' > \sum_{i=1}^{n=100} d_{i\,min}$, Then no new medoids and repeat step 7 for M_n. If all elements in M_n are case 2, then repeat step 7 for M_{n+1}.
3. After processing step 7 for all medoids, repeat step 7 for M_1 until all the elments in each medoids are case 2
4. Stop

The computational cost is $O(k(n-k)^2)$, where n is number of data and k is number of clusters.

2.2 Clustering with unknown number of clusters k

2.2.1 New Clustering validation method

Clustering validation, which examine the correctness of clustering results by analyzing the quality of some specific characteristics of the geometric structure of the results with indices, is a common way to find the desired number of clusters k. There are two main sorts of it, external and internal clustering validation. As the external one needs prior information in the data, only internal clustering validation is adopted in this project. In general, people evaluate the results by using an internal index such as Calinski-Harabasz and silhouette which is matched to clustering method they used. Then this index will quantify the quality of each clustering result with number of cluster k_n. After optimization, the solution k will be solved. Most of the solution k is either global maximum or minimum.
In this project, new clustering validation method is adopted. First, by using all the internal indices which are related to the clustering method, to calculate the value of each clustering result with number of cluster k_n. Then, for each index, find all the k_n while its index value is the local maximum or minimum. The optimal solution k is the mode.
In PAM clustering, it is related to inter and intra-cluster variance. Calculating the minimum distance between each point and medoids is dealing with inter-cluster problem. The swapping phase is dealing with intra-cluster problem. CH, DB(Davies-Bouldin), SD, S_Dbw, Friedman, McClain can evaluate the intra and inter-cluster distance. The common indices are CH, DB, SD, S_Dbw. In the paper Understanding of Internal Clustering Validation Measures [6], it claims that S_Dbw is the best index compare with the three indices mentioned above. However, its algorithm returns error when applying the data of this project in clv R package. As I am not a CS student and dont have enough experience in programming, S_Dbw is not adopted in this report. For SD, the distance function, which is 2-norm formula not squared 2-norm, cannot be modified in clv package. Hence, only CH and DB are adopted in this project.

2.2.2 Calinski-Harabasz index

$$CH(k_n) = \frac{SS_B}{SS_W} \times \frac{N - k_n}{k_n - 1} \tag{5}$$

where $SS_B = \sum_{i=1}^{k} n_l \|c_l - c\|^2$, measuring the inter-cluster variance based on 2-norm formula and $SS_W = \sum_{i=1}^{k} \sum_{s \in M_l} \|x - c_l\|^2$, measuring the intra-cluster variance based on 2-norm

formula. N is the number of data. n_l is the number of data in cluster l. c_l is the centroid of cluster l. c is the mean of the all data. x is a data point.
The calculation is processed through the clustering.evaluation in MATLAB. The distance formula, 2-norm, is fixed. It is not the squared 2-norm, the distance formula used in PAM clustering. Therefore, it might not give accurate result. However, it is more common that SD. So, CH is used. All the local maximums are found.

2.2.3 Davies-Bouldin index

The general formula of DB:

$$DB = \frac{1}{k_n} \sum_{i=1}^{k_n} \max_{i \neq j} \left\{ \frac{dima(c_i) + dima(c_j)}{dist(c_i, c_j)} \right\} \tag{6}$$

where i means i^{th} cluster, dima(c_i) and dist(c_i, c_j) are the intra- and inter-cluster diameters respectively which can be defined in several ways.
By considering PAM-clustering, dima(c_i) is complete diameter (i.e. the distance between two the most remote objects belonging to the same cluster c_i) and dist(c_i, c_j) is single linkage distance (i.e. the closet distance between two samples belonging to two different clusters c_i and c_j) and calculated with 1-norm. There are the reasons.
First, PAM is squared 2-norm. Due to the limitation of the clv package, the best option of the distance function is 1-norm, largest for all p-norm. In DB, the desired value is minimum. Therefore, in order to find it easily, we have to maximize the value for the number of clusters which is wrong. As a result, we have to maximize dima(c_i) and minimize dist(c_i, c_j). All the local minimum knee points are found.

3 Discussion

3.1 With known no. of cluster k

The data set is separated into 4 parts which involve 10 people randomly in each part. The input of the algorithm is one part each time. Therefore, the result will be represented by mean. Given the number of clusters k = 10, the result are shown in Table 1.

	LKS	Gauss	This project
Accuracy	-	0.8770	0.9810
Recall	0.8555	0.3850	0.9050
Specificity	-	0.9317	0.9894
Precision	-	0.6660	0.9185
F1_Score	-	0.4120	0.8991

Table 1: The result of SIFT PAM Clustering with new kernel function (mean) on ORL (10 subjects) with given number of clusters k = 10.

Remarks: LKS is the method used in the paper written by P.Ji, H.li, M.Salzmann[7]. Gauss is similar to the method in this project, but the kernel function is Gaussian kernel function with sigma = 16.5239 is the standard deviation of the 2-norm of the data subset.

It shows that the method used in this project is the best for every evaluation metrics.

3.2 With unknown no. of cluster k

New approach in clustering validation is adopted to find the optimal k by input k from 2 to 99. Unfortunately, the result (table 2) is bad, which cannot find the unique solution in every part. Two of them doesnt include 10 which is the real k. There are several reasons. The formulas of CH and modified DB are based on statistic concept. However, the output is integer. Hence, the result will easily 'shift' to another integer which is incorrect. Apart from that, CH and modified DB do not fit the situation of the algorithm in this project because CH and DB is not calculated with squared 2-norm. The distance functions are 2-norm and 1-norm respectively. Also, due to time constraint, I cannot use more indices which are suitable for PAM. Otherwise, the solution may be unique. However, in one another data subset which has the same size as the test data introduced above, the optimal solution is 10, 100% correct. It is considered as over-fitted result. In short, the effectiveness of this new clustering validation method needs further research.

Part 1	Part 2	Part 3	Part 4
8	26	7	4
10	39	10	44
13	46	39	48
21	51	42	55
40	56	47	57
44	58	50	59
54	60	63	67
64	-	67	73
67	-	77	-
74	-	-	-

Table 2: Optimal solution k on each part of ORL with new clustering validation.

4 Conclusion

In clustering with known number of cluster, the SIFT new kernel function PAM have quite good performance compared with a paper. Also, it is suitable for object classification. With unknown number k, the result of new approach is bad. There are several reasons, one of it is the clustering validation indices are not well modified, which cannot match PAM perfectly. Unfortunately, due to time limitation, it didn't fixed. Hence, the effectiveness of this validation method needs further research. Nowadays, the research trend focus on deep learning as it's potential is larger. First, the general process of applying it involves training which is supervised learning. Also, some algorithms such as Alexnet, 7 layers Deep CNN, can further process the data which acts as a function like SIFT. Moreover, deep learning can finds more than one kernel functions automatically. To conclude, unsupervised learning is difficult task and clustering is less popular.

References

[1] Steinhaus, Hugo. "Sur La Division Des Corps Matriels En Parties. (French)." Bull. Acad. Pol. Sci., Cl. III 4 (1957): 801-04.

[2] Blum, Manuel, Jost Tobias Springenberg, Jan Wulfing, and Martin Riedmiller. "A Learned Feature Descriptor for Object Recognition in RGB-D Data." 2012 IEEE International Conference on Robotics and Automation, 2012. doi:10.1109/icra.2012.6225188.

[3] The Database of Faces. Accessed May 04, 2019.
https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[4] Li, Stan Z. Handbook of Face Recognition. 2011.

[5] Lindeberg, Tony. ”Scale Invariant Feature Transform.” Scholarpedia 7, no. 5 (2012): 10491. doi:10.4249/scholarpedia.10491.

[6] Liu, Yanchi, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. ”Understanding of Internal Clustering Validation Measures.” 2010 IEEE International Conference on Data Mining, 2010. doi:10.1109/icdm.2010.35.

[7] Pan, Ji, Reid Ian, Garg Ravi, Li Hongdong, and Salzmann Mathieu. ”Adaptive Low-Rank Kernel Subspace Clustering.” July 16, 2007.