An online weight-based clustering algorithm for faces selection

Qianmu Li[1, a], Dayi Huang[1,b]

[1]School of Computer Science and Engineering, NUST, Nanjing 210094, China

[a]liqianmu@126.com, [b]seasonhdy@outlook.com

Keywords: online;weight;cluster;faces selection.

Abstract. With the scale of the application of face recognition getting bigger and bigger, A thinking of selecting faces as a preprocessing has conducted on how to improve the system's efficiency. After summarized and analyzed such a thinking, this paper has discussed the pros and cons of face data clustering by using the traditional K-means clustering algorithm. In addition, this paper has further proposed an online weight-based clustering algorithm (OWCA for short) that strengthens the advantages of face data clustering, restricts the computational complexity and develops an online processing scheme. Experimental results on the ORL face library demonstrated the correctness and effectiveness of OWCA.

Introduction

The main purpose of this paper is to propose a kind of process of selecting faces from a face set, which is called 'online weight-based clustering algorithm'. It applies the clustering thought to face classification, considers the weights as the basis for selecting the cluster-center and provides an online processing scheme.

In real life, we often come across with some problems in the scope of face selection. For example, in the scenario of an online security system based on face recognition, we can use the face selection as a prefiltering of the faces which collected by camera over a period of time, it could reduce the network load and the computing pressure of the server[1]; In the scenario of a face database updating, we should select faces those are correct, good quality and non-redundant as the basis of modification to the database[2], etc.These problems can be described as such a process: select a subset as small as possible from a face set, where the subset should represent the original face set well enough.

With researches into this problem, we found that if we use k-means clustering algorithm(KCA for short)[3] to cluster the face set, then select the corresponding face of the cluster-center from the results, in the ideal state, the set of selected faces is the subset that we want. But in practice, KCA had encountered with lots of difficulties. With this in mind, this paper proposes a process named online weight-based clustering algorithm(OWCA for short), which retains the advantages of KCA and avoids its disadvantages, restricts the scale of calculation and provides a kind of online processing scheme. After that, the authors has proceeded theoretical discussions through 2D point set and demonstrated the correctness and effectiveness of OWCA through a experiment on the real face set in the end.

Face selection problem

Some definitions. There are some definitions to note in this paper:

(1) The distance between two faces: A float number in the range of $[0,1]$, calculated by face recognition algorithm. $Dist(a,b)$ is used as the distance of face a and face b.

(2) The number of people that a face set is covering(NPC for short): In a face set, if all faces are captured from n people, that can be considered as the face set covering n people.

(3) The ideal distribution state(IDS for short): The cluster is conformed to the globular shape, and data samples are aggregative within the cluster, scattered among cluster relatively.

(4) The representativeness of a face: In a face set S, the representativeness of the face a equals to

$$\frac{1}{\sum_{b \neq a} Dist(a,b)}, b \in S. \quad (1)$$

(5) The special face: In a face set S, the special face, from the perspective of human, will be considered as the face without a similar company, or the face x that $\min\{Dist(x,y), y \in S\} > \mu$ (where μ is a threshold) from the perspective of mathematics.

(6) The redundant face: In a face set that all faces is similar to each other, if one face of it has been selected, the other faces will be considered as redundant faces.

Problem descriptions. From this paper, the face selection can be described as such a process: To select a subset as small as possible from a face set, where the subset should represent the original face set well enough. More generally, a subset can represent the original face set if and only if it assures:

(1) The NPC of the subset equals to the original face set's.

(2) Each face of the subset has the maximized representativeness for its cluster.

(3) All the special faces has been selected into subset.

Particularly, In the scenario of a growing face set:

(1) Provide an online scheme.

(2) The efficiency achieving the real-time.

Clustering on the face set

Let's starts with a brief introduction of K-means clustering algorithm(KCA for short), so from the very beginning, to discuss the feasibility and benefit of applying its clustering thought to the face set. KCA, presented by J.B, MacQueen(1967), is one of the most classical and widely used clustering algorithm based on similarity(distance for math). This kind of algorithm classifies objects together if they were similar, and its result is usually being a cluster set where clusters are compact and independent[3].

The discussion of k-means clustering algorithm on the face set. For the purposes of illustration, this paper assuming a 2D point set to be a face set, and the distance among faces will be considering as the Euclidean distance among points. As shown in A of Fig. 1, there is a point set which contains 23 points that under the IDS:

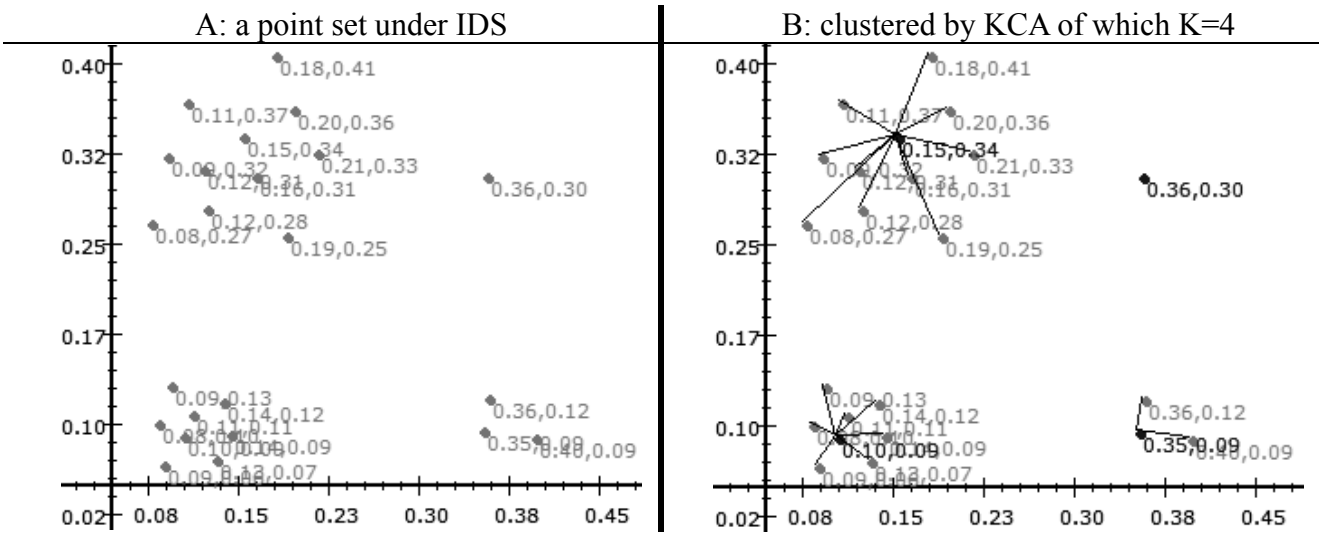

Figure 1: KCA clustering on the point set under IDS

With this in mind, the authors obtain the result shown in B of Fig. 1 by using a KCA of which K=4, where black points are the points nearest a clustering center, each gray points will be linked to the cluster-center of its cluster with a solid line. What if we select all the black points as a subset of the point set?

(1) Redundant filting: input with 23 points, output with 4 points, decrease the size, no cluster missing.
(2) Special faces are selected: noises and outliers will belong to a cluster of their own.
(3) Select properly: the selected points, also known as the cluster-center, possess the error-resistant ability as equalization point. In this basis we can agree that they can represent the cluster that they belong to more or less.
However, all of these are hypothesises under IDS, it is rarely happens in real life:
(1) The results can only divided into k clusters. As shown in A of Fig. 2.
(2) It is easily affected by the initial points and discrete points. As shown in B and C of Fig. 2.
(3) The points nearest a cluster-center may not possess the maximized representativeness. On the 2D point set, the cluster-center of KCA its coordinates are $x=\frac{1}{n}\sum_{i=1}^{n}x_i, y=\frac{1}{n}\sum_{i=1}^{n}y_i$, whichs could only assure the Error variance function

$$J(x,y)=\sum_{i=1}^{n}[(x-x_i)^2+(y-y_i)^2] \tag{2}$$

reaches its minimum value, since it could not assure the Distances sum function

$$D(x,y)=\sum_{i=1}^{n}\sqrt{(x-x_i)^2+(y-y_i)^2} \tag{3}$$

does as it should have for the maximized representativeness, fermat point is a special case[4].
(4) The calculation of KCA is offline, we should give the whole face set at the beginning;
(5) The scale of calculation increases rapidly with the number of points.

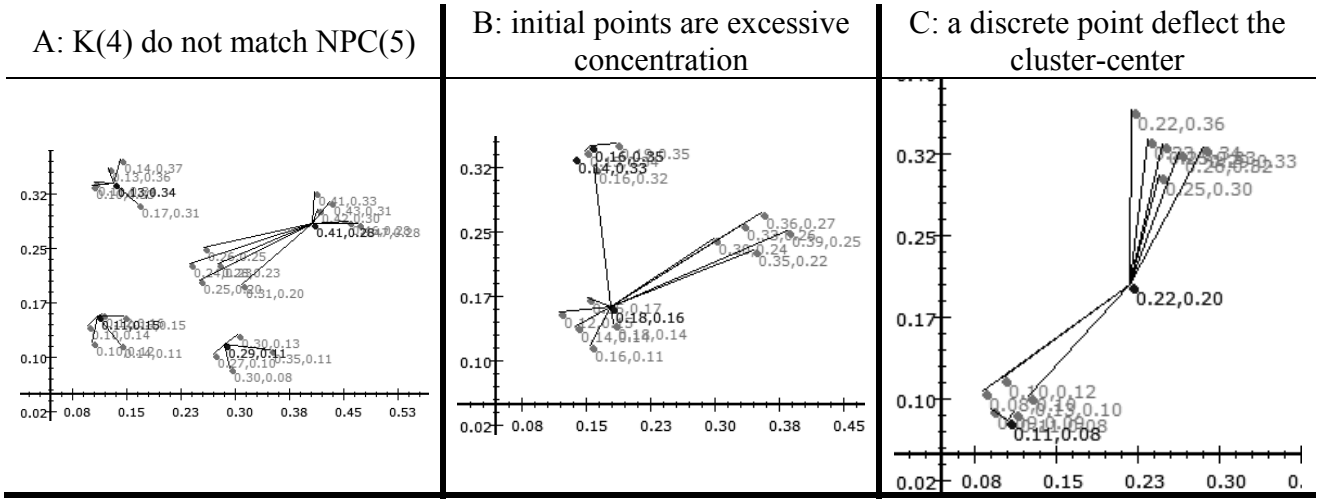

Figure 2: KCA In real life

Online weight-based clustering algorithm. KCA, as previously stated, can not satisfied the face set clustering in real life, but it is also being a regret to give such a thinking up. With researches, this paper has further proposed an online weight-based clustering algorithm(OWCA for short) that strengthens the advantages of face data clustering, restricts the computational complexity and develops an online processing scheme. the workflow is as follows:

Table 1: the workflow of online weight-based clustering algorithm

1. Given the block size B, the distance threshold μ, an empty cluster set S. Use n as the number of clusters in S, m as the number of current data samples, set $n=0, m=0$.

1) Use S_i as the i-th cluster of S; for any $s \in S$, use $s.center$ as the cluster-center of s, use $s.size$ as the number of data samples in s;

2) Every existing data sample belongs and only belongs to one cluster, use s_i as the i-th data sample of s, while using w_i as the weight of s_i in the scope of s.

2. For a new data sample x, to look for an index I, which satisfied

$$Dist(x,(S_I).center) = \min\{Dist(x,(S_i).center), i=1,\cdots,n\} \text{ and } Dist(x,(S_I).center) \le \mu \qquad (4)$$

at the same time. On the basis of this:

1) if such an I exists, then classify x into S_I, and set $m=m+1, (S_I).size+=1$;

2) if not exists, create a new cluster and add it into S, then classify x into this cluster, and set $I=n+1, n=n+1, m=m+1, (S_I).size+=1$.

3. Update cluster $s=S_I$ as below:

1) compute $d_{x,s_i} = Dist(x,s_i), i=1,\cdots,s.size-1$;

2) set $w_i = w_i + d_{x,s_i}, i=1,\cdots,s.size-1$;

3) set $w_{s.size} = \sum_{i=1}^{s.size-1} d_{x,s_i}$;

4) look for an index J, which satisfied $w_J = \min\{w_i, i=1,\cdots,s.size\}$, then set $s.center = s_J$;

4. To judge the number of current data samples:

1) if $m < B$, then go to step 2.

2) if $m \ge B$, outputs S as the block result, clear S and set $n=0, m=0$, then goto step 2.

OWCA begins its iterative process automatically after the initialization as the step 1. listed in Table 1. During its processing procedure, it awaits and accepts new input data sample from outside continually, for every new data sample, OWCA classify them and update the cluster set, output will take place when the number of all data samples in OWCA reaches its upper limit.

The discussion of online weight-based clustering algorithm on the face set. The discussion is happening under a 2D point set as well. It is easy to learn from the workflow above that OWCA still possesses the abilities of 'Redundant filting', 'Special faces are selected' and 'Select properly' described as before. And, there are more to it than that:

(1) No limitations on the number of cluster.

(2) No initial cluster-centers worrying.

(3) Discrete points do not affect the result under a proper threshold.

(4) The cluster-center has the maximized representativeness. For the cluster-centers of OWCA's result, whichs have the minimum weight to its cluster, according to the Eq. 1 and the step 3 of the workflow shown in Table 1, they represent their respective clusters best.

(5) Provide an online scheme.

(6) The calculation times have an upper limit. In the worst case, that all points will be classified together into one cluster, its calculation times is:

$$[(B-1)+1+2+3+4+\cdots+(B-1)]\frac{c}{B} = \frac{1}{2}(B-1)(B+2)\frac{c}{B}. \qquad (5)$$

where the first $B-1$ in the square brackets is the times that calculate the distances between the new point and every cluster-center, while the rest is the times of calculating the distances between the new

point and every other point of its cluster. $\frac{c}{B}$ is the times of block-calculating, where c is the total number of points, B is the block size.

Experiments

In earlier discussions the authors use a 2D point set as the face set for the purpose of demonstration, visual but rationalistic. This section will be utilized to test the OWCA on the face set in real life. In order to discuss more generally, this paper uses face recognition cloud service Face++ API[5] as the standard of calculating the distance between two faces, the result will be referred directly, the details of the implementation will not be covered here.
First and foremostl, this paper has established some evaluation criterions for clustering on a face set:

Table 2: evaluation criterions for clustering on a face set

	Content
1	the number of faces on (input \ output)
2	NPC of (input \ output)
3	calculation times(upper limit \ statistics from program)
4	the elapsed time of one calcultion on average[s]
5	the number of special faces on (input \ output)
6	special faces hit ratios[%]
7	the number of mis-classification

For OWCA, the input will be the set of faces those are passed to OWCA, the output will be the set of faces corresponding to cluster-centers on the cluster result. Moreover, there are some criterions deserved a bit of attention:
(1) The upper limit of calculation times: It can be computed directly by Eq. 5 with B and c.
(2) The special faces on input: They are the faces those of artificial selected from input face set.
(3) The special faces on output: Some clusters in the cluster result of OWCA may only contains one face, all these faces are considered as the special faces on output.
(4) Special faces hit ratios: use S_a as the set of special faces on ouput, S_m as the set of special faces on input, $Count(S)$ as the number of faces on S. on these basises the special faces hit ratios can be computed as:

$$\frac{Count(S_a \cap S_m)}{Count(S_a)} \times 100\% \tag{6}$$

(5) The number of mis-classification: the number of faces which does not match its cluster-center as the same person in the cluster result.

Experiment on the ORL face set. The ORL database contains 400 images, 10 for each person. A face samples of s1 is as below:

Figure 3: ORL s1 face samples

Simply to see that a face samples could contain standard faces like sample_000.jpg, and special faces like sample_004.jpg, sample_009.jpg. there is not a definition of special face in ORL database, so we need to select artificially. Number all the 400 pictures into 000 to 399 and select two special faces for each person:

Table 3: special faces of artificial selection

person	*special faces*	*person*	*special faces*	*person*	*special faces*	*person*	*special faces*
s1	001、009	s11	106、107	s21	204、208	s31	302、308
s2	011、019	s12	113、119	s22	216、219	s32	310、312
s3	025、029	s13	121、126	s23	220、226	s33	322、326
s4	032、036	s14	130、138	s24	233、238	s34	336、339
s5	044、046	s15	146、149	s25	244、249	s35	343、349
s6	057、058	s16	150、151	s26	256、257	s36	351、359
s7	062、067	s17	165、168	s27	260、269	s37	362、369
s8	075、077	s18	172、178	s28	277、279	s38	371、372
s9	084、085	s19	183、187	s29	280、285	s39	387、389
s10	094、095	s20	192、199	s30	295、297	s40	391、397

There are 10 face samples of each person on the face set, so it is better to set the block size B equals to a multiple of 10 to avoid block-across calculating which may affect the statistics of some criterions; also, the block size B should divide 400 evenly, otherwise there will be some faces are detained in the memory after the final output. The authors set B to 20 here.

For the distance threshold μ , by calculating same-person distances of the first 5 people, the author found that distances between the same people's faces are close to 0.11 probably. The authors set μ to 0.11 here.

The evaluation of an OWCA which use 20 as B ,0.11 as μ on the ORL face set is as below:

Table 4: evaluations for clustering on the ORL face set of OWCA

	Content	*value*
1	the number of faces on (input \ output)	400\97
2	NPC of (input \ output)	40\40
3	calculation times(upper limit \ statistics)	4180\2172
4	the elapsed time of one calcultion on average[s]	0.2704
5	the number of special faces on (input \ output)	80\22
6	special faces hit ratios[%]	81.8
7	the number of mis-classification	0

With the study of this evaluation, the authors found that:
(1) Reduce the redundancy greatly. The amount of data samples curtailed from 400 to 97, the authors also calculated the average distance between different people's faces is near 0.267 on the ORL face set, which means that the distances threshold can be set larger to 0.2. This will reduce redundancy furthermore.
(2) The correctness of the results. The number of mis-classification is 0, which also known as, every cluster of the result contains only one person's face samples, and hence avoids the omission of people.
(3) The scale of calculation is acceptable. In this experiment, the statistical calculation times is almost the half of its upper limit, which also means that, in general cases, the intensity of calculation is acceptable. The average elapsed time of one calculation is much greater than it really is, because of the uncertain network environment. It falls to 0.0160s for the local face matching program.
(4) The selection of special faces. For the number of special faces on (input \ output), a background factor should be mentioned is that the selection artificially was restricted to two of each person, ignoring the fact that there may be no obvious special faces in a person's face data samples. Moreover, the artificially selection always depends on the lighting conditions and face poses, while Face++ eliminating these factors before calculating as normalizing. The special faces hit ratios shows that the selection of special faces by OWCA still acceptable.

Conclusion

In this paper, we address the face selection problem and find that in most cases we should select a subset as small as possible and the subset can be as good as possible representing the original face. The authors are inspired by k-means clustering algorithm(KCA for short), after clustering by KCA on the face set, we can get the most pleasant surprise to find that the result has three advantages: redundant filting, selecting properly, special faces are selected. However, with further researches, KCA has been proved that it is not suitable for face selecting directly. In our paper, we present a novel algorithm which called online weight clustering algorithm (OWCA for short). It can not only maintain the above three advantages, but also fit the face set scenes. We test the correctness and effectiveness of OWCA by ORL face set. Experiments show that the proposed algorithm produces better results compared with the conventional algorithm.

References

[1] Soyata, T; Muraleedharan, R; Minseok Kwon; et al. Cloud-Vision: Real-time face recognition using a mobile-cloudlet-cloud acceleration architecture[C]. Computers and Communications (ISCC), 2012 IEEE Symposium on , pp.59-66, 1-4 July 2012

[2] Pavani, Sri-Kaushik; Sukno, Federico M.; Butakoff, Constantine; et al. A confidence-based update rule for self-updating human face recognition systems[C]. Lecture Notes in Computer Science, v 5558 LNCS, p 151-160, 2009.

[3] J. MacQueen. SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS[C]. 5-th Berkeley Symposium on Mathematical Statistics and Probability,1967:281~297.

[4] S Gueron, R Tessler. The fermat-steiner problem[J]. American Mathematical Monthly,2002.

[5] Megvii Inc. Face++ Research Toolkit. www.faceplusplus.com, December 2013.