Determination of Number of Clusters in *K*-Means Clustering and Application in Colour Image Segmentation

Siddheswar Ray and Rose H. Turi

School of Computer Science and Software Engineering
Monash University, Wellington Road, Clayton, Victoria, 3168, Australia
E-mail: {sid,roset}@csse.monash.edu.au

Abstract:

The main disadvantage of the k-means algorithm is that the number of clusters, *K*, must be supplied as a parameter. In this paper we present a simple validity measure based on the intra-cluster and inter-cluster distance measures which allows the number of clusters to be determined automatically. The basic procedure involves producing all the segmented images for 2 clusters up to K_{max} clusters, where K_{max} represents an upper limit on the number of clusters. Then our validity measure is calculated to determine which is the best clustering by finding the minimum value for our measure. The validity measure is tested for synthetic images for which the number of clusters in known, and is also implemented for natural images.

Keywords: Clustering; *K*-means; Colour image segmentation; Intra-cluster distance; Inter-class distance.

1. Introduction

Many approaches to image segmentation have been proposed over the years [1-12]. Of these various methods, clustering is one of the simplest, and has been widely used in segmentation of grey level images [13-15]. Techniques such as k-means [16], isodata [16], and fuzzy c-means [17,18] have been around for quite a while, however, their application to colour images has been limited. Although colour images have increased dimensionality by requiring three bands such as red, green and blue, clustering techniques can be easily extended to cope with this. The k-means and fuzzy c-means algorithms require the number of clusters to be known beforehand, and the isodata algorithm has six parameters which must be supplied by the user. In order to supply the information required by the aforementioned algorithms, the user must have some knowledge about the image, and this may not be the case. The new method is based on the k-means algorithm and it overcomes the limitation of having to indicate the number of clusters by incorporating a validity measure based on the intra-cluster and inter-cluster distance measures. The performance of our proposed cluster validity measure is compared with the results obtained using the Davies-Bouldin index[19] and Dunn's index [20].

2. K-means Method

The k-means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described by Tou and Gonzalez [16].

1. Choose *K* initial cluster centres $z_1(1)$, $z_2(1)$, ,..., $z_K(1)$.

2. At the *k*-th iterative step, distribute the samples {x} among the *K* clusters using the relation,

$$x \in C_j(k) \text{ if } \left\| x - z_j(k) \right\| < \left\| x - z_i(k) \right\|$$

for all $i = 1, 2, \ldots, K;\ i \neq j;$ where $C_j(k)$ denotes the set of samples whose cluster centre is $z_j(k)$.

3. Compute the new cluster centres $z_j(k+1)$, $j = 1, 2, \ldots, K$ such that the sum of the squared distances from all points in $C_j(k)$ to the new cluster centre is minimized. The measure which minimizes this is simply the sample mean of $C_j(k)$. Therefore, the new cluster centre is given by

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in C_j(k)} x, \quad j = 1,2,...,K$$

where N_j is the number of samples in $C_j(k)$.

4. If $z_j(k+1) = z_j(k)$ for $j = 1, 2, \ldots, K$ then the algorithm has converged and the procedure is terminated.

Otherwise go to Step 2.

It is obvious in this description that the final clustering will depend on the initial cluster centres chosen and on the value of K. The latter is of the most concern since this requires some prior knowledge of the number of clusters present in the data, which, in practice, is highly unlikely.

3. Cluster Validity Measures

3.1 Existing Measures

Many criteria have been developed for determining cluster validity [19-25], all of which have a common goal to find the clustering which results in compact clusters which are well separated. The Davies-Bouldin index [19], for example, is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The objective is to minimize this measure as we want to minimize the within-cluster scatter and maximize the between-cluster separation. Bezdek and Pal [21] have given a generalization of Dunn's index [20]. Also, by considering five different measures of distance function between clusters and three different measures of cluster diameter, they obtained fifteen different values of the generalized Dunn's index. For details on expressions for these indices please see [21]. Let us denote these fifteen indices by D_{ij} where $1 \leq j \leq 5$ and $1 \leq i \leq 3$. D_{11} represents the original definition of Dunn's index.

3.2 Proposed Measure

Since the k-means method aims to minimize the sum of squared distances from all points to their cluster centres, this should result in compact clusters. We can therefore use the distances of the points from their cluster centre to determine whether the clusters are compact. For this purpose, we use the intra-cluster distance measure, which is simply the distance between a point and its cluster centre and we take the average of all of these distances, defined as

$$\text{intra} = \frac{1}{N} \sum_{i=1}^{K} \sum_{x \in C_i} \left\| x - z_i \right\|^2$$

where N is the number of pixels in the image, K is the number of clusters, and z_i is the cluster centre of cluster C_i. We obviously want to minimize this measure. We can also measure the inter-cluster distance, or the distance between clusters, which we want to be as big as possible. We calculate this as the distance between cluster centres, and take the minimum of this value, defined as

$$\text{inter} = \min(\left\| z_i - z_j \right\|^2), \; i = 1,2,...,K-1, \quad j = i+1,...,K$$

We take only the minimum of this value as we want the smallest of this distance to be maximized, and the other larger values will automatically be bigger than this value.

Since we want both of these measures to help us determine if we have a good clustering, we must combine them in some way. The obvious way is to take the ratio, defined as:

$$validity = \frac{\text{intra}}{\text{inter}}$$

Since we want to minimize the intra-cluster distance and this measure is in the numerator, we consequently want to minimize the validity measure. We also want to maximize the inter-cluster distance measure, and since this is in the denominator, we again want to minimize the validity measure. Therefore, the clustering which gives a minimum value for the validity measure will tell us what the ideal value of K is in the k-means procedure.

4. Description of Method

A number of colour spaces exist in which the segmentation of images can be performed [26]. However, the method discussed here and the results reported later are all based on the use of the (red, green, blue) colour space. This method, however, could be easily implemented in any colour space.

We basically want to produce the segmented images for 2 up to K_{max} clusters, where K_{max} is an upper limit on the number of clusters, and then calculate the validity measure to determine which is the best clustering, and, therefore, what is the optimal value of K. We do this by first forming one cluster containing all the pixels in the image. Then an iterative process begins where, unless the number of clusters is equal to K_{max}, the cluster having maximum variance is split into two. Once the cluster is split, we make use of the k-means procedure to obtain the clustering for this new number of clusters. Once all the clusters have been formed, the validity measure can be calculated for each of them to determine what the optimal value of K is.

Since the k-means algorithm aims to minimize the average intra-cluster distance, it is most likely that the cluster having maximum variance will be separated by the k-means procedure when the number of clusters is increased. Therefore, when we require the number of clusters to be increased, we split the cluster having maximum variance, so the k-means procedure is given good starting cluster centres. We calculate the variance of the three components for cluster C_I as

$$\sigma_{ij}^2 = \frac{1}{N_i}\sum_{x \in C_i}(x - z_{ij})^2, \quad i = 1,2,...,K \qquad j = 1,2,3.$$

where N_i is the number of pixels in cluster C_i and x is the vector representing each pixel's red, green and blue components as x_1, x_2, and x_3, respectively. This gives us three variance values, but we ultimately want just one value, which we can use to compare the variance of each cluster. We take the average variance of the three components by adding them up and dividing the sum by 3. This gives us the following variance values

$$\sigma_i^2 = \frac{1}{3}\sum_{j=1}^{3}\sigma_{ij}^2 \qquad i = 1,2,...,K$$

When splitting a cluster, we take into account all three of the red, green and blue components. Given the cluster C_i whose cluster centre is z_i, we wish to obtain two new cluster centres z_i' and z_i''. We split cluster C_i by creating two new values for each component, which are centred around the cluster centre value. The two new cluster centres are calculated as

$$z_i' = (z_{i1} - a_1,\ z_{i2} - a_2,\ z_{i3} - a_3)$$

$$z_i'' = (z_{i1} + a_1,\ z_{i2} + a_2,\ z_{31} + a_3)$$

where a_1, a_2 and a_3, are constants. The values for these constants are determined by taking into account the minimum and maximum values for each colour component occurring in the cluster. The constants, a_j will be the values which are half of the smaller of $(z_{ij} - min_j)$ and

$(z_{ij} - max_j)$, where min_j is the minimum value for the j-th colour component and max_j is the maximum value for the j-th component. This results in the two new cluster centres being well separated, but also still well within the original cluster.

With the above method we could use any validity measure such as the Davies-Bouldin index or Dunn's index. For Dunn's index we would want to find the clustering which maximizes this index.

5. Experimental Results

Experiments were conducted for both synthetic images and natural images. In the following two sub-sections, namely, 5.1 and 5.2, results obtained for these images are discussed. Primarily due to space limitations, the input images and the values of the proposed and existing validity criteria are not presented. They will be shown during the presentation of the paper at the conference.

5.1 Synthetic Images

This method was first implemented with synthetic images for which the ideal clustering was known beforehand. A total of five synthetic images were used with varying numbers of clusters. The first two images, *syn1* and *syn2*, have four clusters. Three of the clusters had uniformly distributed values with a range of 30 in one colour component each, and the other cluster had a constant value. *syn1* has clusters with varying sizes, while *syn2* had equal sized clusters. The third synthetic image, *syn3* has nine clusters each of the same size and each having values uniformly distributed with a range of 30 in one colour component. *syn4* has 16 clusters of equal size, 15 of which have values uniformly distributed with a range of 30 in one colour component and one cluster of constant value. Finally *syn5* has only two clusters of equal size each of which has values uniformly distributed with a range of 30 in one colour component.

The algorithm described in the previous section was executed on each of these images with K_{max} set to 25, as the ideal cluster numbers are obviously well below this. It was found that the minimum values of the validity measure occurred at the correct number of clusters for each of the synthetic images. The Davies-Bouldin and Dunn's indexes were both implemented for these synthetic images which also resulted in their optimum values occurring at the correct number of clusters.

5.2 Natural Images

The next step involved testing this method with real images. A total of eleven images were selected which represent a wide variety of colour images from the segmentation point of view. These images are called *balls*, *Lenna*, *molecule*, *teapot*, *ant*, *blond*, *jet*, *mandrill*, *peppers*, *mouse* and *rose*. Once again K_{max} was set to 25 because this is suitably large enough given that from a visual point of view we would not identify more than 25 colour regions in these images.

The first thing we noticed was that there was a tendency for the minimum value of the validity measure to occur for small numbers of clusters in the range of 2, 3, or 4. This is due to a large inter-cluster distance value occurring when the number of clusters is this low, resulting in the validity measure being very small. The only exception occurred for the image *molecule* for which the minimum value for the validity measure was produced for 8 clusters. In general, we expect that colour images will have a number of clusters greater than 2, 3 or 4. So, instead of simply selecting the clustering which leads to the minimum value of the validity measure, we look for the first local maximum in the validity measure, where a local maximum is defined to occur at k if

$$validity(k-1) < validity(k) > validity(k+1),$$

$$k \geq 3.$$

Once we find the first local maximum in the validity measure, occurring for k clusters, we then select the smallest value of the validity measure between $k + 1$ and K_{max} clusters. By this definition the first possible local maximum can occur for three clusters, so the smallest number of clusters which could be selected is four. In general we do not expect that colour images will have only 2 or 3 clusters. By applying the above stated modified rule for the synthetic images, the correct number of clusters can still be found, except for *syn2* which only had 2 clusters.

The number of clusters produced (1) based on simply the global minimum and (2) by following the modified rule for our proposed measure are shown in Table 1. The most common failing with the resulting segmentation based on global minimum of the validity measure was that that part of the objects were classified together with the background. This was caused due to an inadequate number of clusters to represent the regions in the images. The modified rule of first finding the local maximum overcomes this problem, as this ensures that such low numbers of clusters can never be selected. Segmentation results based on the modified rule show a vast improvement.

Table 1: Number of Clusters

Image	Global Minimum	Modified Minimum
balls	2	10
Lenna	2	11
molecule	8	8
teapot	3	8
ant	3	6
blond	2	5
jet	2	5
mandrill	4	10
peppers	2	6
mouse	2	6
rose	3	6

The Davies-Bouldin and Dunn's indexes also have a tendency to select a small number of clusters. Therefore, a similar modified rule can be applied to them. The Davies-Bouldin index resulted in a reasonable number of clusters for some images, however, for *balls*, *Lenna*, and *mouse* the number of clusters obtained by the modified rule are too small to adequately represent the regions in these images.

Our validity measure worked more consistently for the natural images than either the Davies-Bouldin index or Dunn's indices. The number of clusters produced by our measure produced good segmentation results for each of the natural images as opposed to the other measures for which a good segmentation could not be found for each of the images.

6. Conclusion

By incorporating the validity measure based on the intra-cluster and inter-cluster distance measures, the number of clusters present in an image can be determined automatically. The validity measure proposed here works well with synthetic images, producing a minimum value for the expected number of clusters. Although there is a tendency to select smaller cluster numbers for natural images, this is due to the inter-cluster distance being much greater and greatly affecting the validity measure. We overcome this by looking for a local maximum for the validity measure and then by finding the minimum value after the local maximum. By using this modified rule, the smallest number of clusters that can be selected is four. This is not a real problem because natural colour images can be expected to have more than two or three clusters. The modified rule still allows the optimal number of clusters to be selected for the synthetic images, except for the image with only two clusters as two clusters cannot be selected by this modified rule. The Davies-Bouldin index and Dunn's indexes could not detect the correct number of clusters for all the natural images, however, our validity measure

performed more consistently for all of the natural images, producing good segmentation results.

As minor modifications to this algorithm, we could use median cluster centre representation instead of mean cluster centre representation, and we could use absolute distance instead of Euclidean distance to calculate the distance between a pixel and its cluster centre or between cluster centres. We could also use any of the number of different colour spaces available. As an improvement we could also incorporate the context of the image, as a given pixel is expected to be highly correlated with its neighbouring pixels values. This could be achieved by taking into account the values of the neighbouring pixels.

This method is not restricted to colour images. It can be easily extended to cope with any dimensionality, so this method may also be used for multispectral images. Similarly, there is no reason why this method cannot be used for grey scale images, which have only one dimension.

References

[1] N.R. Pal and S.K. Pal, A review on image segmentation techniques, *Pattern Recognition*, vol. 26, pp. 1277-1294, 1993.

[2] K.S. Fu and J.K. Mui, A survey on image segmentation, *Pattern Recognition*, vol. 13, pp. 3-16, 1981.

[3] R.M. Haralick and L.G. Shapiro, Survey image segmentation techniques, *Comput. Vision Graphics Image Process*., vol. 29, pp. 100-132, 1985.

[4] A. Rosenfeld and L.S. Davis, Image segmentation and image models, *Proc. IEEE*, vol. 67, pp. 764-772, 1979.

[5] P.K. Sahoo, S. Soltani, A.K.C. Wong and Y.C. Chen, A survey of thresholding techniques, *Comput. Vision Graphics Image Process*., vol. 41, pp. 233-260, 1988.

[6] A. Perez and R.C. Gonzalez, An iterative thresholding algorithm for image segmentation, *IEEE Trans. Pattern Anal. Machine Intell*., vol. 9, pp. 742-751, 1987.

[7] T. Peli and D. Malah, A study of edge detection algorithms, *Comput. Graphics Image Process*., vol. 20, pp. 1-21, 1982.

[8] R. Ohlander, K. Price and D.R. Reddy, Picture segmentation using a recursive region splitting method, *Comput. Graphics Image Process*., vol. 8, pp. 313-333, 1978.

[9] S.L. Horowitz and T. Pavlidis, Picture segmentation by directed split and merge procedure, *Proc. 2nd Int. Joint Conf. Pattern Recognition*, pp. 424-433, 1974.

[10] J. Liu and Y. Yang, Multiresolution color image segmentation, *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 689-700, 1994.

[11] M. Amadasun and R.A. King, Low level segmentation of multispectural images via agglomerative clustering of uniform neighbours, *Pattern Recognition*, vol. 21, pp. 261-268, 1988.

[12] B. Bhanu and B.A. Rarvin, Segmentation of natural scene, *Pattern Recognition*, vol. 20, pp. 487-496, 1987.

[13] G.B. Coleman and H.C. Andrews, Image segmentation by clustering, *Proc. IEEE*, vol. 67, pp. 773-785, 1979.

[14] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, New Jersey: Prentice Hall, 1988.

[15] R. Nevatia, Image segmentation, In T.Y. Young and K.S. Fu (Eds.), *Handbook of Pattern Recognition and Image Processing*, Orlando: Academic Press, 1986.

[16] J.T. Tou and R.C. Gonzalez, *Pattern Recognition Principles*, Massachusetts: Addison-Wesley, 1974.

[17] M.M. Trivedi and J.C. Bezdek, Low-level segmentation of aerial images with fuzzy clustering, *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-16, pp. 589-598, 1986.

[18] Y.W. Lim and S.U. Lee, On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques, *Pattern Recognition*, vol. 23, pp. 935-952, 1990.

[19] D.L. Davies and D.W. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Anal. Machine Intell.* Vol. 1, pp. 224-227, 1979.

[20] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybern.*,vol. 3,pp. 32-57,1973.

[21] J.C. Bezdek and N.R. Pal, Some new indexes of cluster validity, *IEEE Trans. Syst. Man. Cybern*., vol. 28, pp. 301-315, 1998.

[22] G.W. Milligan, Clustering validation: Results and implications for applied analyses, In P. Arabie, L.J. Hubert and G. De Soete (Eds.), *Clustering and Classification*, Singapore: World Scientific, pp. 341-375, 1996.

[23] G.W. Milligan and M.C. Cooper, An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, vol. 50, pp. 159-179, 1985.

[24] M.C. Cooper and G.W. Milligan, The effect of measurement error on determining the number of clusters in cluster analysis, In W. Gaul and M. Schader (Eds.), *Data, Expert Knowledge and Decisions*, Berlin: Springer-Verlag, pp. 319-328, 1988.

[25] N.R. Pal and J.C. Bezdek, On cluster validity for the fuzzy c-means model, *IEEE Trans. Fuzzy Systems*, vol. 3, pp. 370-379, 1995.

[26] Y. Ohta, T. Kanade and T. Sakai, Color information for region segmentation, *Comput. Graphics Image Process.*, vol. 13, pp. 222-241, 1980.