Selection of *K* in *K*-means clustering

D T Pham*, **S S Dimov,** and **C D Nguyen**
Manufacturing Engineering Centre, Cardiff University, Cardiff, UK

The manuscript was received on 26 May 2004 and was accepted after revision for publication on 27 September 2004.

DOI: 10.1243/095440605X8298

Abstract: The *K*-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the requirement for the number of clusters, *K*, to be specified before the algorithm is applied. This paper first reviews existing methods for selecting the number of clusters for the algorithm. Factors that affect this selection are then discussed and a new measure to assist the selection is proposed. The paper concludes with an analysis of the results of using the proposed measure to determine the number of clusters for the *K*-means algorithm for different data sets.

Keywords: clustering, *K*-means algorithm, cluster number selection

1 INTRODUCTION

Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing. Data clustering has many engineering applications including the identification of part families for cellular manufacture.

The *K*-means algorithm is a popular data-clustering algorithm. To use it requires the number of clusters in the data to be pre-specified. Finding the appropriate number of clusters for a given data set is generally a trial-and-error process made more difficult by the subjective nature of deciding what constitutes 'correct' clustering [**1**].

This paper proposes a method based on information obtained during the *K*-means clustering operation itself to select the number of clusters, *K*. The method employs an objective evaluation measure to suggest suitable values for *K*, thus avoiding the need for trial and error.

The remainder of the paper consists of five sections. Section 2 reviews the main known methods for selecting *K*. Section 3 analyses the factors influencing the selection of *K*. Section 4 describes the proposed evaluation measure. Section 5 presents the results of applying the proposed measure to select *K* for different data sets. Section 6 concludes the paper.

**Corresponding author: Manufacturing Engineering Centre, Cardiff University, Cardiff CF24 OYF, UK.*

2 SELECTION OF THE NUMBER OF CLUSTERS AND CLUSTERING VALIDITY ASSESSMENT

This section reviews existing methods for selecting *K* for the *K*-means algorithm and the corresponding clustering validation techniques.

2.1 Values of *K* specified within a range or set

The performance of a clustering algorithm may be affected by the chosen value of *K*. Therefore, instead of using a single predefined *K*, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the specific characteristics of the data sets. At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering.

Reported studies [**2**–**18**] on *K*-means clustering and its applications usually do not contain any explanation or justification for selecting particular values for *K*. Table 1 lists the numbers of clusters and objects and the corresponding data sets used in those studies. Two observations could be made when analysing the data in the table. First, a number of researchers [**5**–**7**, **9**] used only one or two values for *K*. Second, several other researchers [**1**, **3**, **11**, **13**, **16**] utilized relatively large *K* values compared with the number of objects. These two actions contravene the above-mentioned guidelines for selecting *K*. Therefore, the clustering results do not always correctly represent the performance of the tested algorithms.

Table 1 The number of clusters used in different studies of the *K*-means algorithm

Reference	Numbers of clusters *K*	Number of objects *N*	Maximum *K*/*N* ratio (%)
[**2**]	32, 64, 128, 256, 512, 1024	8 192	12.50
	32, 64, 128, 256, 512, 1024	29 000	
	256	2 048	
[**3**]	600, 700, 800, 900, 1000	10 000	10.00
	600, 700, 800, 900, 1000	50 000	
[**4**]	4, 16, 64, 100, 128	100 000	0.13
	4, 16, 64, 100, 128	120 000	
	4, 16, 64, 100, 128	256 000	
[**5**]	4	564	0.70
	4	720	
	4	1 000	
	4	1 008	
	4	1 010	
	4	1 202	
	4	2 000	
	4	2 324	
	4	3 005	
	4	4 000	
	4	6 272	
	4	7 561	
[**6**]	6	150	4.00
[**7**]	10	2 310	0.43
	25	12 902	
[**8**]	2, 4, 8	Not reported	Not reported
[**9**]	2, 4	500	3.33
	2, 4	50 000	
	2, 4	100 000	
	10	300	
[**10**]	1, 2, 3, 4	10 000	0.04
[**11**]	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	500	20.00
[**12**]	100	10 000	2.00
	50	2 500	
[**13**]	7	42	16.66
	1, 2, 3, 4, 5, 6, 7	120	
[**14**]	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	250	5.60
[**15**]	8, 20, 50, 64, 256	10 000	2.56
[**16**]	5000	50 000	50.00
	5000	100 000	
	5000	200 000	
	5000	300 000	
	5000	433 208	
	100	100 000	
	250	200 000	
	1000	100 000	
	1000	200 000	
	1000	300 000	
	1000	433 208	
	40	20 000	
	10, 20, 30, 40, 50, 60, 70, 80	30 000	
	50, 500, 5000	10 000	
	50, 500, 5000	50 000	
	50, 500, 5000	100 000	
	50, 500, 5000	200 000	
	50, 500, 5000	300 000	
	50, 500, 5000	433 208	
[**17**]	250	80 000	10.00
	250	90 000	
	250	100 000	
	250	110 000	
	250	120 000	
	50, 100, 400	4 000	
	50, 100, 400	36 000	
	250	80 000	
	250	90 000	
	250	100 000	
	250	110 000	
	250	120 000	
	50, 100, 150	4 000	
	50, 100, 150	36 000	
	50	800 000	
	500	800 000	
[**18**]	3, 4	150	6.67
	4, 5	75	
	2, 7, 10	214	

(*continued*)

In general, the performance of any new version of the *K*-means algorithm could be verified by comparing it with its predecessors on the same criteria. In particular, the sum of cluster distortions is usually employed as such a performance indicator [**3**, **6**, **13**, **16**, **18**]. Thus, the comparison is considered fair because the same model and criterion are used for the performance analysis.

2.2 Values of *K* specified by the user

The *K*-means algorithm implementation in many data-mining or data analysis software packages [**19**–**22**] requires the number of clusters to be specified by the user. To find a satisfactory clustering result, usually, a number of iterations are needed where the user executes the algorithm with different values of *K*. The validity of the clustering result is assessed only visually without applying any formal performance measures. With this approach, it is difficult for users to evaluate the clustering result for multi-dimensional data sets.

2.3 Values of *K* determined in a later processing step

When *K*-means clustering is used as a pre-processing tool, the number of clusters is determined by the specific requirements of the main processing algorithm [**13**]. No attention is paid to the effect of the clustering results on the performance of this algorithm. In such applications, the *K*-means algorithm is employed just as a 'black box' without validation of the clustering result.

2.4 Values of *K* equated to the number of generators

Synthetic data sets, which are used for testing algorithms, are often created by a set of normal or uniform distribution generators. Then, clustering algorithms are applied to those data sets with the number of clusters equated to the number of generators. It is assumed that any resultant cluster will cover all objects created by a particular generator. Thus, the clustering performance is judged on the basis of the difference between objects covered by a cluster and those created by the corresponding generator. Such a difference can be measured by simply counting objects or calculating the information gain [**7**].

There are drawbacks with this method. The first drawback concerns the stability of the clustering results when there are areas in the object space that contain objects created by different generators. Figure 1a illustrates such a case. The data set shown in this figure has two clusters, A and B, which cover objects generated by generators G_A and G_B respectively. Object X is in an overlapping area between clusters A and B. X has probabilities P_{G_A} and P_{G_B} of being created by G_A and G_B, respectively, and probabilities P_{C_A} and P_{C_B} of being included into clusters A and B, respectively. All four probabilities are larger than 0. Thus, there is a chance for X to be created by generator G_A but covered by cluster B, and vice versa. In such cases, the clustering results will not be perfect. The stability of the clustering results depends on these four probabilities. With an increase in the overlapping areas in the object space, the stability of the clustering results decreases.

The difference between the characteristics of the generators also has an effect on the clustering results. In Fig. 1b where the number of objects of cluster A is five times larger than that of cluster B, the smaller cluster B might be regarded as noise and all objects might be grouped into one cluster. Such a clustering outcome would differ from that obtained by visual inspection.

Unfortunately, this method of selecting *K* cannot be applied to practical problems. The data distribution in practical problems is unknown and also the number of generators cannot be specified.

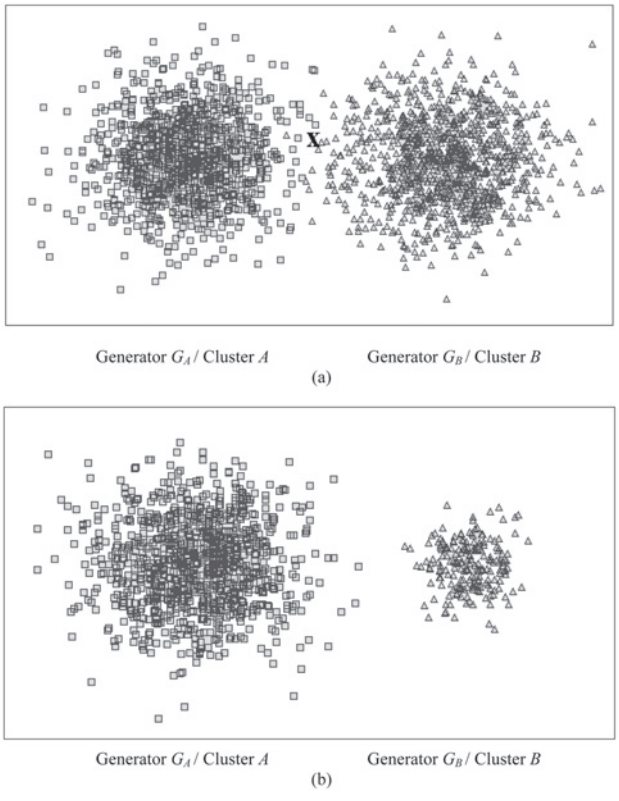


Fig. 1 Effect of the relationship between clusters on the clustering for two object spaces in which (a) an area exists that contains objects created by two different generators and (b) there are no overlapping areas: □, objects generated by G_A; Δ, objects generated by G_B

2.5 Values of *K* determined by statistical measures

There are several statistical measures available for selecting *K*. These measures are often applied in combination with probabilistic clustering approaches. They are calculated with certain assumptions about the underlying distribution of the data. The Bayesian information criterion or Akeike's information criterion [**14**, **17**] is calculated on data sets which are constructed by a set of Gaussian distributions. The measures applied by Hardy [**23**] are based on the assumption that the data set fits the Poisson distribution. Monte Carlo techniques, which are associated with the *null hypothesis*, are used for assessing the clustering results and also for determining the number of clusters [**24**, **25**].

There have been comparisons between probabilistic and partitioning clustering [**7**]. Expectation–maximization (EM) is often recognized as a typical method for probabilistic clustering. Similarly, *K*-means clustering is considered a typical method for partitioning clustering. Although, EM and *K*-means clustering share some common ideas, they are based on different hypotheses, models and criteria. Probabilistic clustering methods do not take into account the distortion inside a cluster, so that a cluster created by applying such methods may not correspond to a cluster in partitioning clustering, and vice versa. Therefore, statistical measures used in probabilistic methods are not applicable in

the *K*-means algorithm. In addition, the assumptions about the underlying distribution cannot be verified on real data sets and therefore cannot be used to obtain statistical measures.

2.6 Values of *K* equated to the number of classes

With this method, the number of clusters is equated to the number of classes in the data sets. A data-clustering algorithm can be used as a classifier by applying it to data sets from which the class attribute is omitted and then assessing the clustering results using the omitted class information [**26**, **27**]. The outcome of the assessment is fed back to the clustering algorithm to improve its performance. In this way, the clustering can be considered to be supervised.

With this method of determining the number of clusters, the assumption is made that the data-clustering method could form clusters, each of which would consist of only objects belonging to one class. Unfortunately, most real problems do not satisfy this assumption.

2.7 Values of *K* determined through visualization

Visual verification is applied widely because of its simplicity and explanation possibilities. Visual examples are often used to illustrate the drawbacks of an algorithm or to present the expected clustering results [**5**, **27**].

The assessment of a clustering result using visualization techniques depends heavily on their implicit nature. The clustering models utilized by some clustering methods may not be appropriate for particular data sets. The data sets in Fig. 2 are illustrations of such cases. The application of visualization techniques implies a data distribution continuity in the expected clusters. If the *K*-means approach is applied to such data sets, there is not

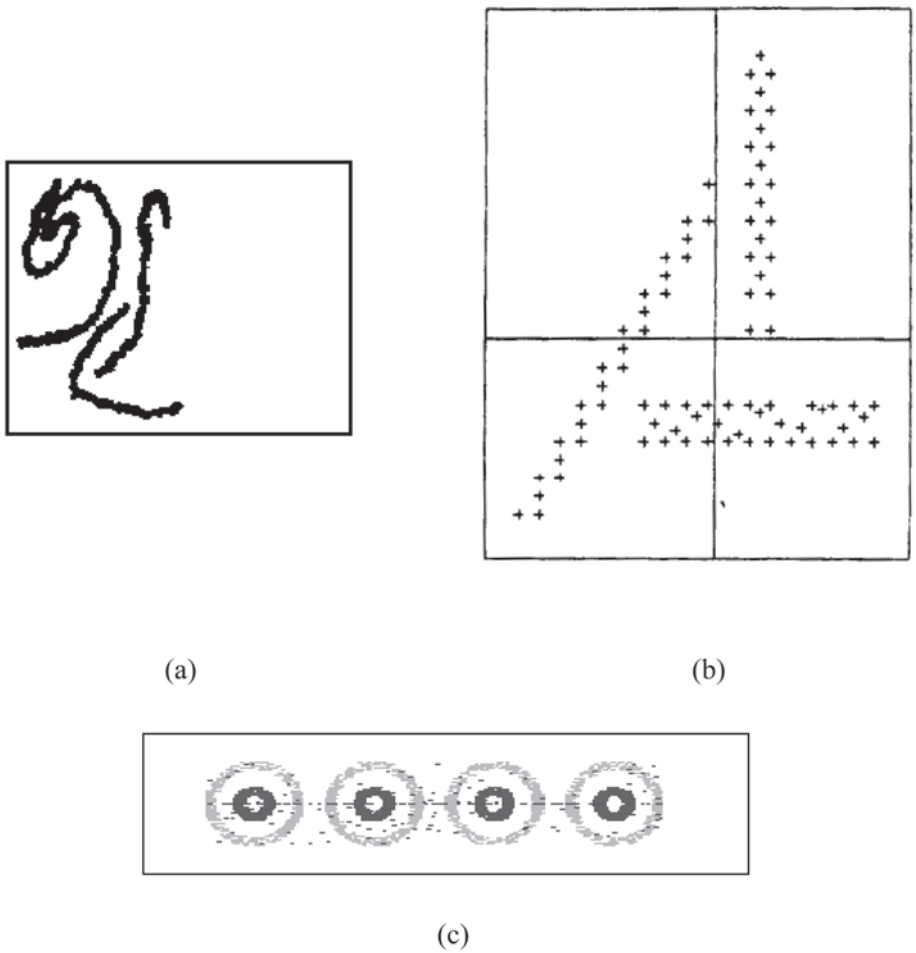


Fig. 2 Data sets inappropriate for the *K*-means approach: (a) data sets with four clusters [**5**]; (b) data sets with three clusters [**23**]; (c) data sets with eight clusters [**27**]. Note that the number of clusters in each data set was specified by the respective researchers

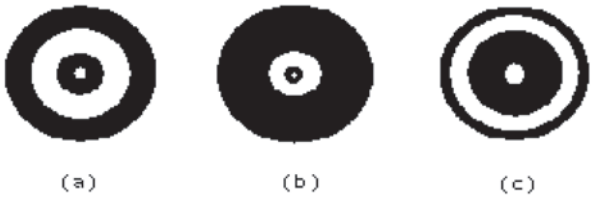


Fig. 3 Variations in the two-ring data set

any cluster that satisfies the *K*-means clustering model and at the same time corresponds to a particular object grouping in the illustrated data sets. Therefore, the *K*-means algorithm cannot produce the expected clustering results. This suggests that the *K*-means approach is unsuitable for such data sets.

The characteristics of the data sets in Fig. 2 (position, shape, size, and object distribution) are implicitly defined. This makes the validation of the clustering results difficult. Any slight changes in the data characteristics may lead to different outcomes. The data set in Fig. 2b is an illustration of such a case. Another example is the series of data sets in Fig. 3. Although two clusters are easily identifiable in the data set in Fig. 3a, the numbers of clusters in the data sets in Figs 3b and c depend on the distance between the rings and the object density of each ring. Usually such parameters are not explicitly defined when a visual check is carried out.

In spite of the above-mentioned deficiencies, visualization of the results is still a useful method of selecting *K* and validating the clustering results when the data sets do not violate the assumptions of the clustering model. In addition, this method is recommended in cases where the expected results could be identified explicitly.

2.8 Values of *K* determined using a neighbourhood measure

A neighbourhood measure could be added to the cost function of the *K*-means algorithm to determine *K* [**26**]. Although this technique has showed promising results for a few data sets, it needs to prove its potential in practical applications. Because the cost function has to be modified, this technique cannot be applied to the original *K*-means algorithm.

3 FACTORS AFFECTING THE SELECTION OF *K*

A function $f(K)$ for evaluating the clustering result could be used to select the number of clusters. Factors that such a function should take into account are discussed in this section.

3.1 Approach bias

The evaluation function should be related closely to the clustering criteria. As mentioned previously, such a relation could prevent adverse effects on the validation process. In particular, in the *K*-means algorithm, the criterion is the minimization of the distortion of clusters, so that the evaluation function should take this parameter into account.

3.2 Level of detail

In general, observers that could see relatively low levels of detail would obtain only an overview of an object. By increasing the level of detail, they could gain *more information* about the observed object but, at the same time, the amount of data that they have to process increases. Because of resource limitations, a high level of detail is normally used only to examine parts of the object [**28**].

Such an approach could be applied in clustering. A data set with n objects could be grouped into any number of clusters between 1 and n, which would correspond to the lowest and the highest levels of detail respectively. By specifying different K values, it is possible to assess the results of grouping objects into various numbers of clusters. From this evaluation, more than one K value could be recommended to users, but the final selection is made by them.

3.3 Internal distribution versus global impact

Clustering is used to find irregularities in the data distribution and to identify regions in which objects are concentrated. However, not every region with a high concentration of objects is considered a cluster. For a region to be identified as a cluster, it is important to analyse not only its internal distribution but also its interdependence with other object groupings in the data set.

In *K*-means clustering, the distortion of a cluster is a function of the data population and the distance between objects and the cluster centre according to

$$I_j = \sum_{t=1}^{N_j} [d(\mathrm{x}_{jt}, \mathrm{w}_j)]^2 \tag{1a}$$

where I_j is the distortion of cluster j, w_j is the centre of cluster j, N_j is the number of objects belonging to cluster j, x_{jt} is the tth object belonging to cluster j, and $d(\mathrm{x}_{jt}, \mathrm{w}_j)$ is the distance between object x_{jt} and the centre w_j of cluster j.

Each cluster is represented by its distortion and its impact on the entire data set is assessed by

its contribution to the sum of all distortions, S_K, given by

$$S_K = \sum_{j=1}^{K} I_j \tag{1b}$$

where K is the specified number of clusters.

Thus, such information is important in assessing whether a particular region in the object space could be considered a cluster.

3.4 Constraints on $f(K)$

The robustness of $f(K)$ is very important. Because this function is based on the result of the clustering algorithm, it is important for this result to vary as little as possible when K remains unchanged. However, one of the main deficiencies of the K-means approach is its dependence on randomness. Thus, the algorithm should yield consistent results so that its performance can be used as a variable in the evaluation function. A new version of the K-means algorithm, namely the incremental K-means algorithm [**29**], satisfies this requirement and can be adopted for this purpose.

The role of $f(K)$ is to reveal trends in the data distribution and therefore it is important to keep it independent of the number of objects. The number of clusters, K, is assumed to be much smaller than the number of objects, N. When K increases, $f(K)$ should converge to some constant value. Then, if, for any intermediate K, $f(K)$ exhibits a special behaviour, such as a minimum or maximum point, that value of K could be taken as the desired number of clusters.

4 NUMBER OF CLUSTERS FOR K-MEANS CLUSTERING

As mentioned in section 3.3, cluster analysis is used to find irregularities in the data distribution. When the data distribution is uniform, there is not any irregularity. Therefore, data sets with uniform distribution could be used to calibrate and verify the clustering result. This approach was applied by Tibshirani *et al.* [**30**]. A data set of the same dimension as the actual data set and with a uniform distribution was generated. The clustering performance on this artificial data set was then compared with the result obtained for the actual data set. A measure known as the '*gap*' statistic [**30**] was employed to assess performance. In this work, instead of generating an artificial data set, the clustering performance for the artificial data set was estimated. Also, instead of the gap statistic, a new and more discriminatory measure was employed for evaluating the clustering result.

When the K-means algorithm is applied to data with a uniform distribution and K is increased by 1, the clusters are likely to change and, in the new positions, the partitions will again be approximately equal in size and their distortions similar to one another. The evaluations carried out in reference [**29**] showed that, when a new cluster is inserted into a cluster ($K = 1$) with a hypercuboid shape and a uniform distribution, the decrease in the sum of distortions is proportional to the original sum of distortions. This conclusion was found to be correct for clustering results obtained with relatively small values of K. In such cases, the sum of distortions after the increase in the number of clusters could be estimated from the current value.

The evaluation function $f(K)$ is defined using the equations

$$f(K) = \begin{cases} 1 & \text{if } K = 1 \\ \dfrac{S_K}{\alpha_K S_{K-1}} & \text{if } S_{K-1} \neq 0, \forall K > 1 \\ 1 & \text{if } S_{K-1} = 0, \forall K > 1 \end{cases} \tag{2}$$

$$\alpha_K = \begin{cases} 1 - \dfrac{3}{4N_{\mathrm{d}}} & \text{if } K = 2 \text{ and } N_{\mathrm{d}} > 1 \end{cases} \tag{3a}$$

$$\alpha_K = \begin{cases} \alpha_{K-1} + \dfrac{1 - \alpha_{K-1}}{6} & \text{if } K > 2 \text{ and } N_{\mathrm{d}} > 1 \end{cases} \tag{3b}$$

where S_K is the sum of the cluster distortions when the number of clusters is K, N_{d} is the number of data set attributes (i.e. the number of dimensions) and α_K is a weight factor. The term $\alpha_K S_{K-1}$ in equation (2) is an estimate of S_K based on S_{K-1} made with the assumption that the data have a uniform distribution. The value of $f(K)$ is the ratio of the real distortion to the estimated distortion and is close to 1 when the data distribution is uniform. When there are areas of concentration in the data distribution, S_K will be less than the estimated value, so that $f(K)$ decreases. The smaller that $f(K)$ is, the more concentrated is the data distribution. Thus, values of K that yield small $f(K)$ can be regarded as giving well-defined clusters.

The weight factor α_K, defined in equation (3), is a positive number less than or equal to 1 and is applied to reduce the effect of dimensions. With $K = 2$, α_K is computed using equation (3a). This equation is derived from equation (7) in reference [**29**], which shows that the decrease in distortion is inversely proportional to the number of dimensions, N_{d}.

As K increases above 2, the decrease in the sum of distortions reduces (the ratio S_K/S_{K-1} approaches 1), as can be seen in Fig. 4. This figure shows the values

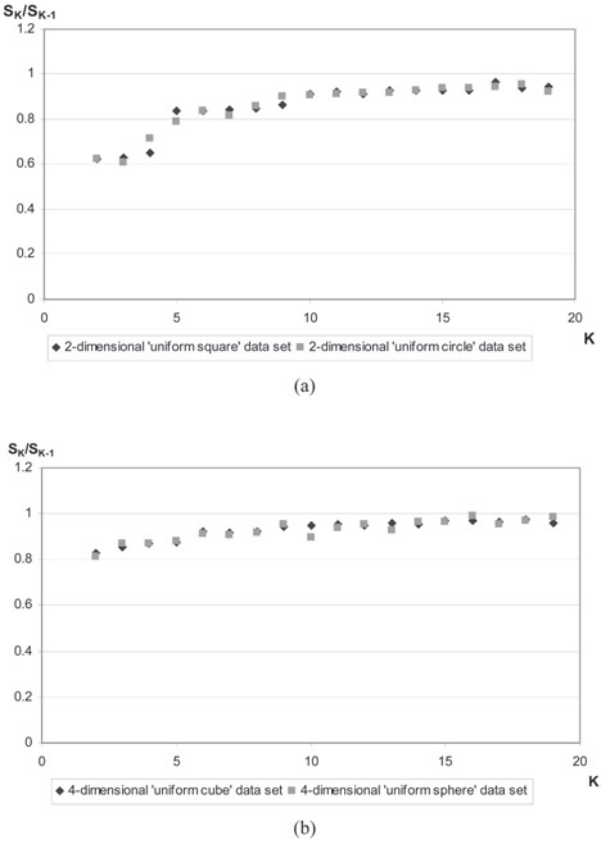


Fig. 4 The ratio S_K/S_{K-1} for data sets having uniform distributions: (a) two-dimensional 'square' and 'circle'; (b) four-dimensional 'cube' and 'sphere'

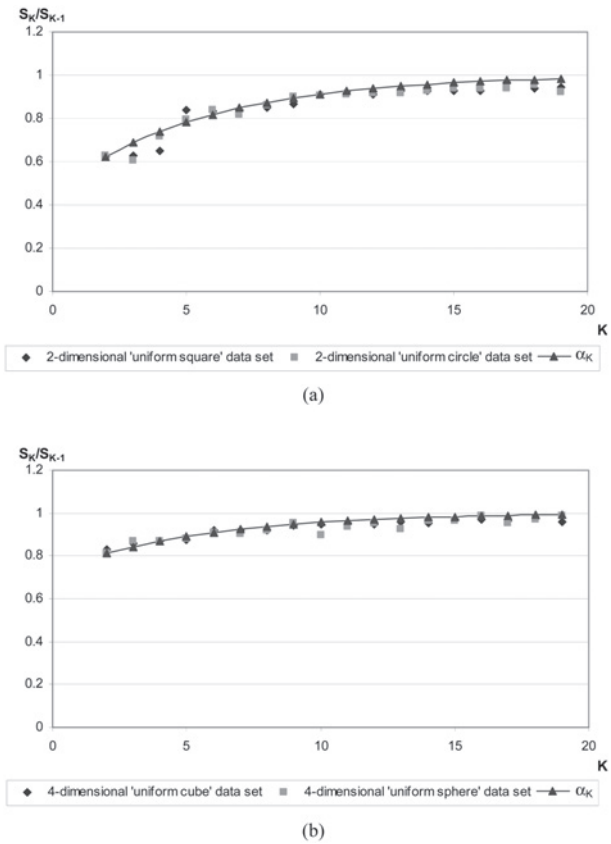


Fig. 5 Comparison of the values of α_K calculated using equation (3b) and the ratio S_K/S_{K-1}

of S_K/S_{K-1} computed for different K when the clustering algorithm is applied to data sets of different dimensions and with uniform distributions. With such data sets, $f(K)$ is expected to be equal to 1 and α_K should be chosen to equate $f(K)$ to 1. From equation (2), α_K should therefore be S_K/S_{K-1} and thus obtainable from Fig. 4. However, for computational simplicity, the recursion equation (3b) has been derived from the data represented in Fig. 4 to calculate α_K. Figure 5 shows that the values of α_K obtained from equation (3b) fit the plots in Fig. 4 closely.

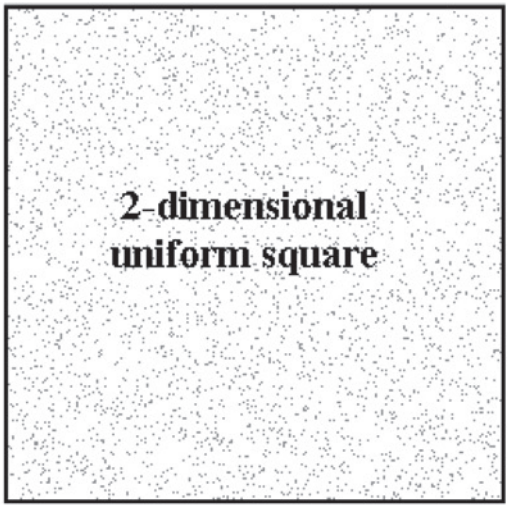
The proposed function $f(K)$ satisfies the constraints mentioned in the previous section. The robustness of $f(K)$ will be verified experimentally in the next section. When the number of objects is doubled or tripled but their distributions are unchanged, the resultant clusters remain in the same position. S_K and S_{K-1} are doubled or tripled correspondingly, so that $f(K)$ stays constant. Therefore, generally, $f(K)$ is independent of the number of objects in the data set.

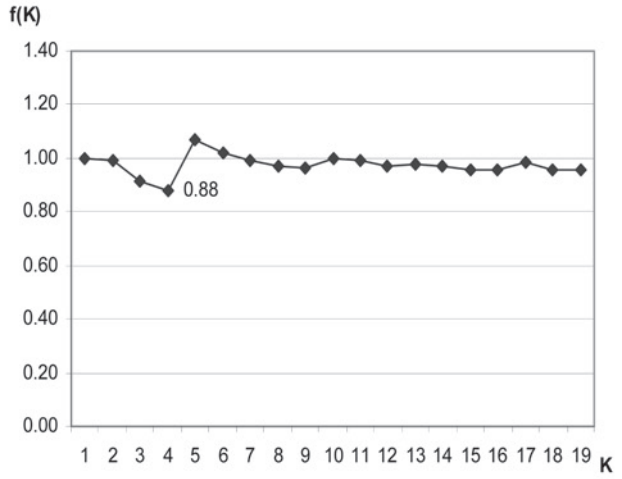
To reduce the effect of differences in the ranges of the attributes, data are normalized before the clustering starts. However, it should be noted that, when the data have well-separated groups of objects, the shape of such regions in the problem space has an effect on the evaluation function. In these cases, the normalization does not influence the local object distribution, because it is a scaling technique that applies to the whole data set.

5 PERFORMANCE

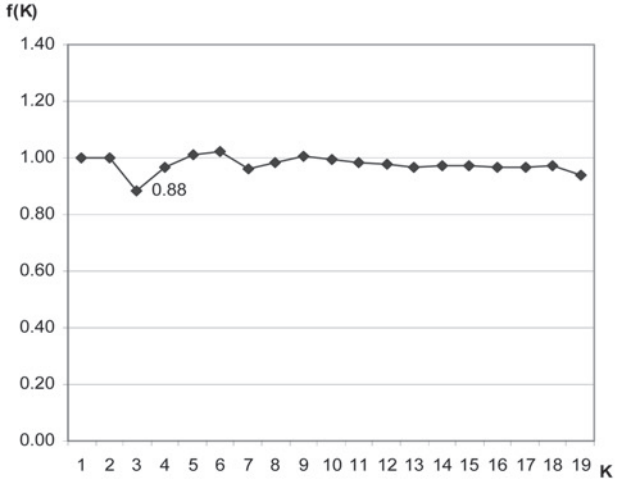
The evaluation function $f(K)$ is tested in a series of experiments on the artificially generated data sets shown in Fig. 6. All data are normalized before the incremental K-means algorithm is applied with K ranging from 1 to 19. $f(K)$ is calculated on the basis of the total distortion of the clusters.

In Figs 6a–c, all objects belong to a single region with a uniform distribution. The graph in Fig. 6a shows that $f(K)$ reflects well the clustering result on this data set with a uniform distribution because $f(K)$ is approximately constant and equal to 1 for all K. When $K = 4$ and $K = 3$ in Figs 6a and b, respectively, $f(K)$ reaches minimum values. This could be attributed to the shape of the areas defined by the objects belonging to these data sets. However, the minimum values of $f(K)$ do not differ significantly

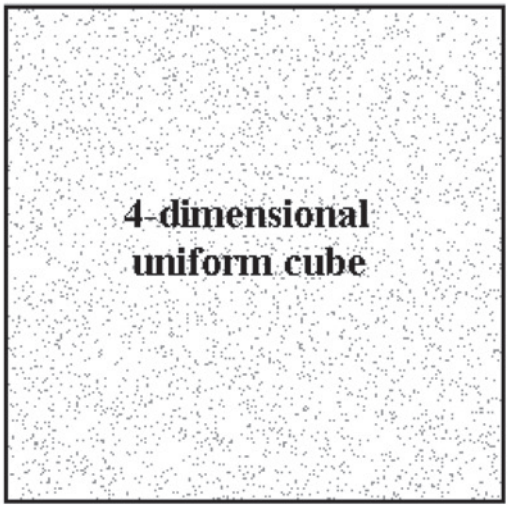


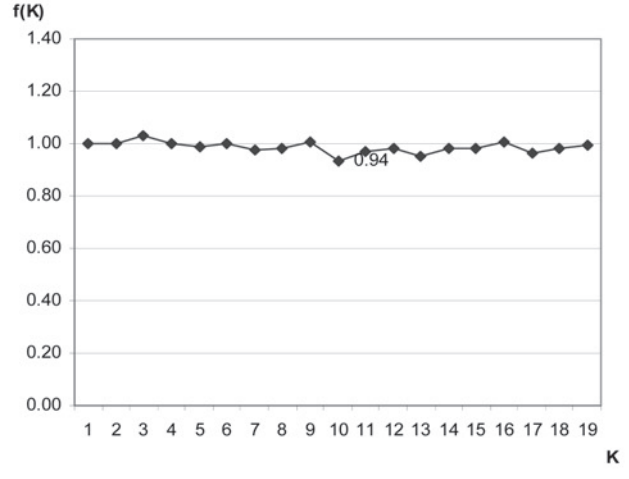


(a)



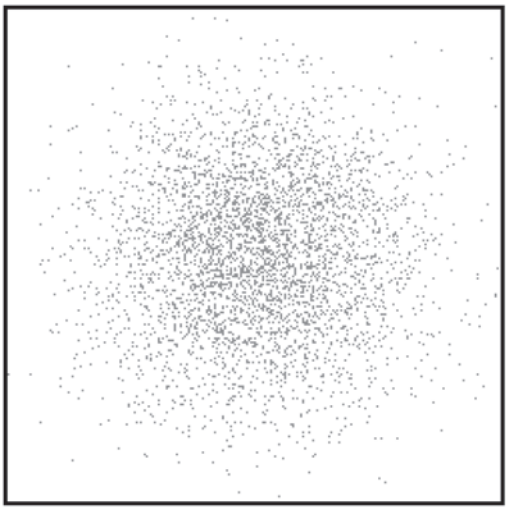
(b)

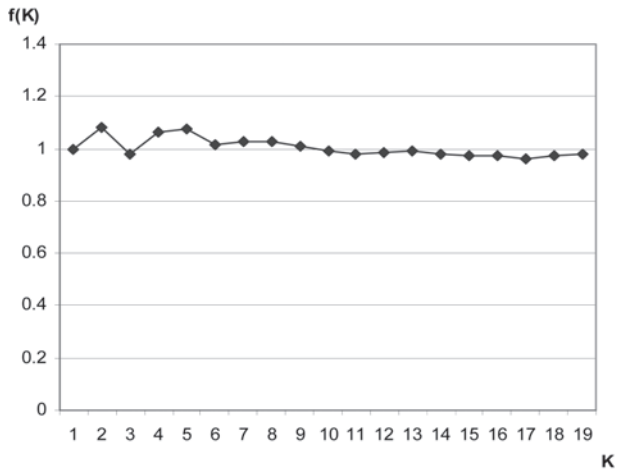




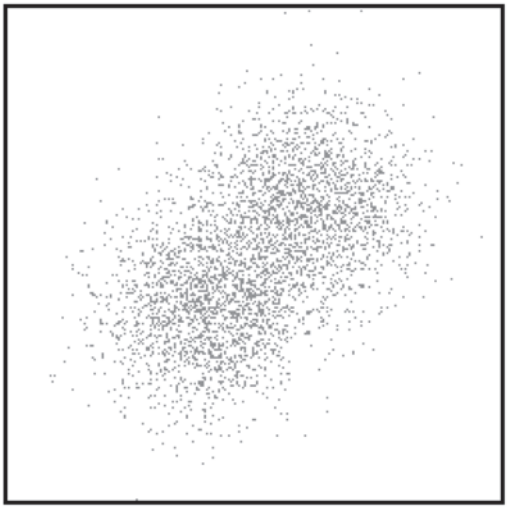
(c)

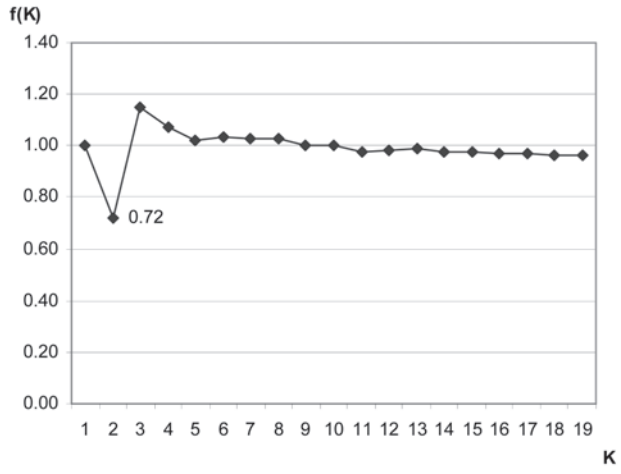
Fig. 6 Data sets and their corresponding $f(K)$



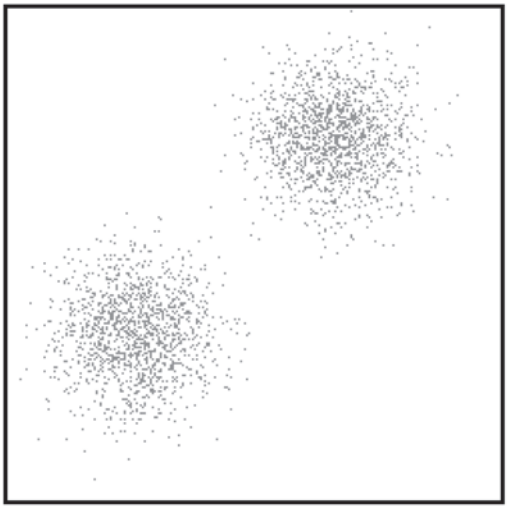


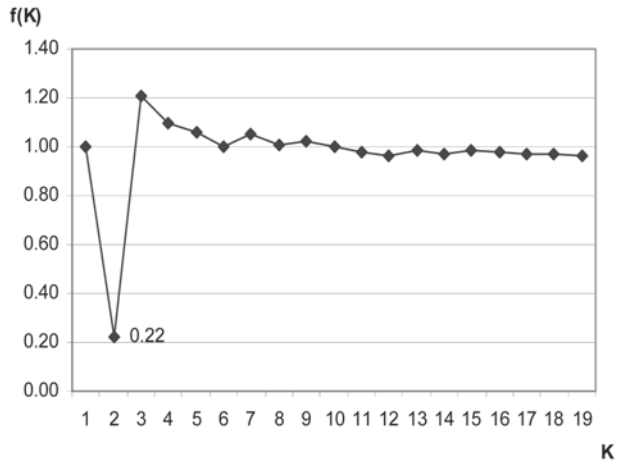
(d)





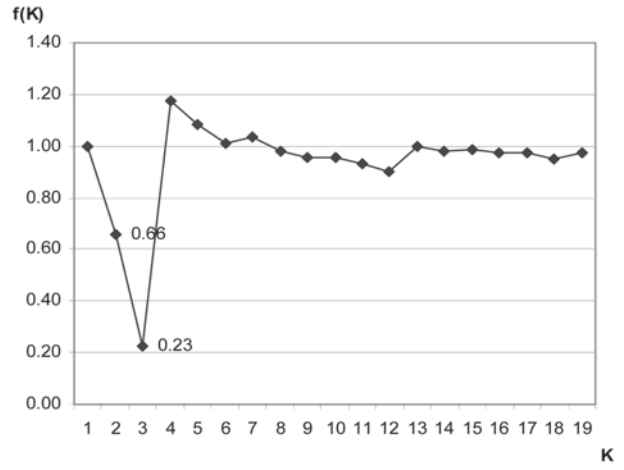
(e)



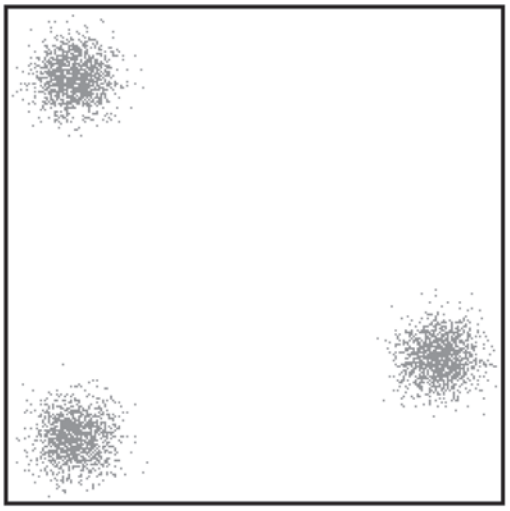


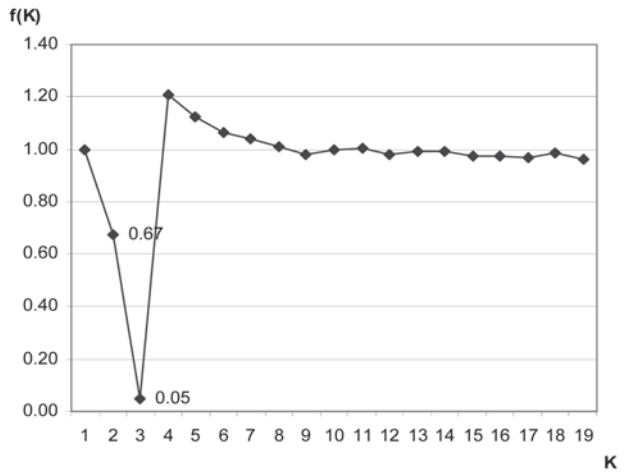
(f)

Fig. 6 Continued

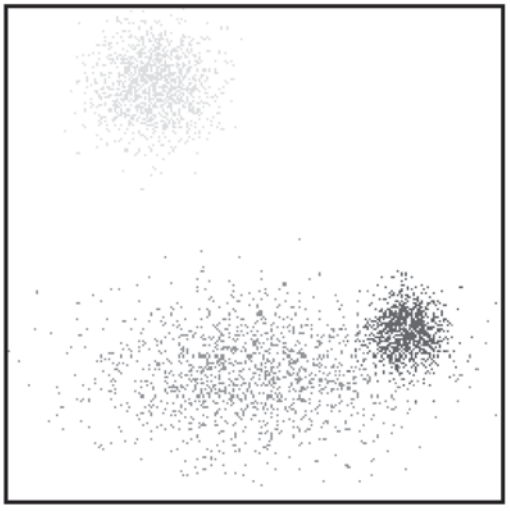


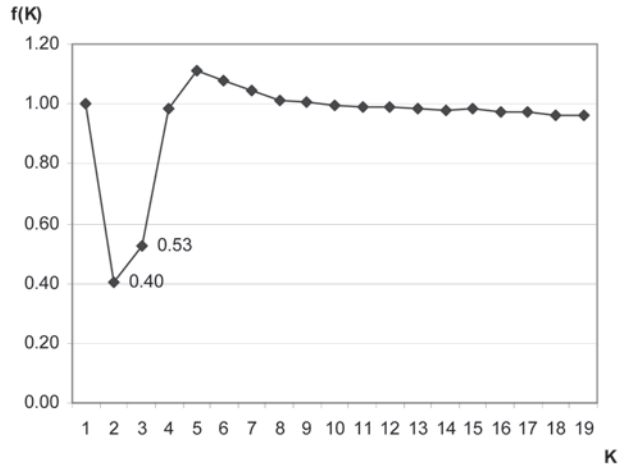
(g)





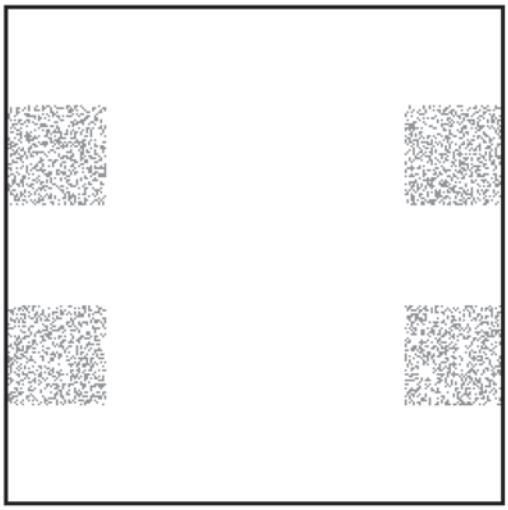
(h)

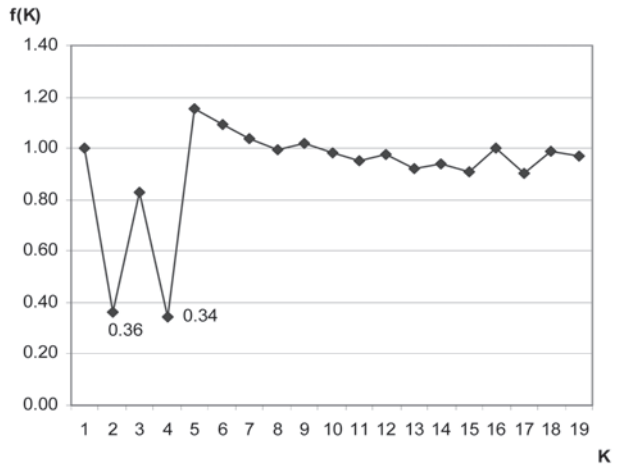




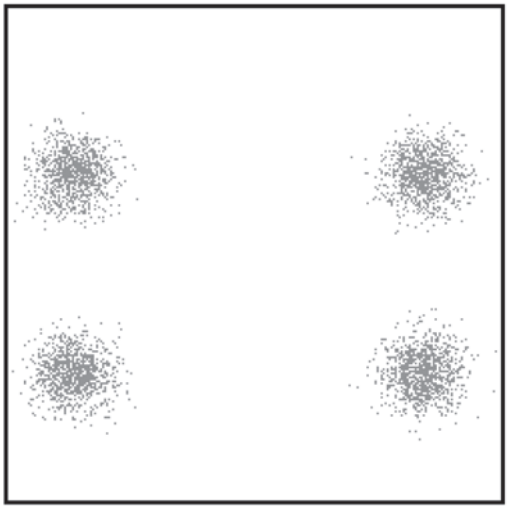
(i)

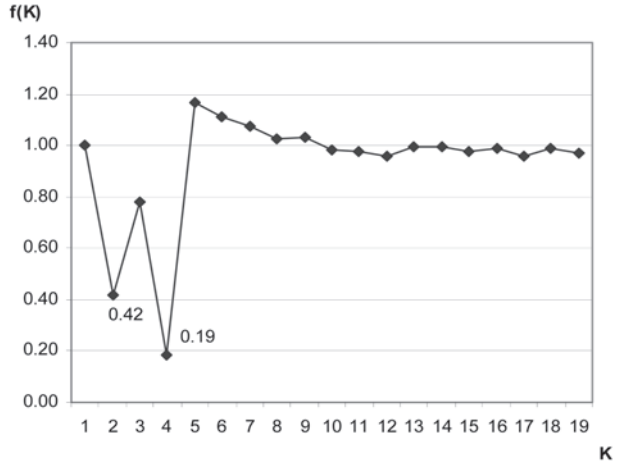
Fig. 6 Continued



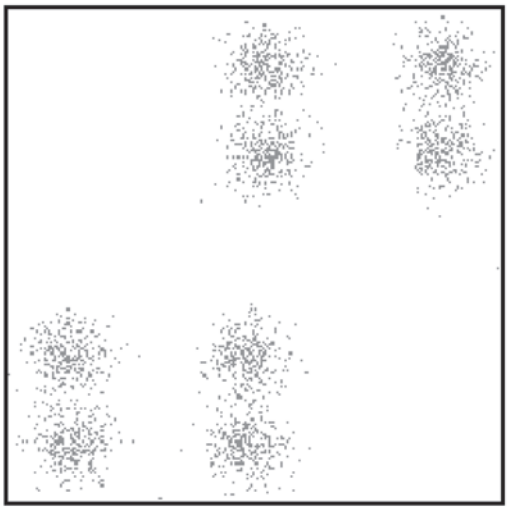


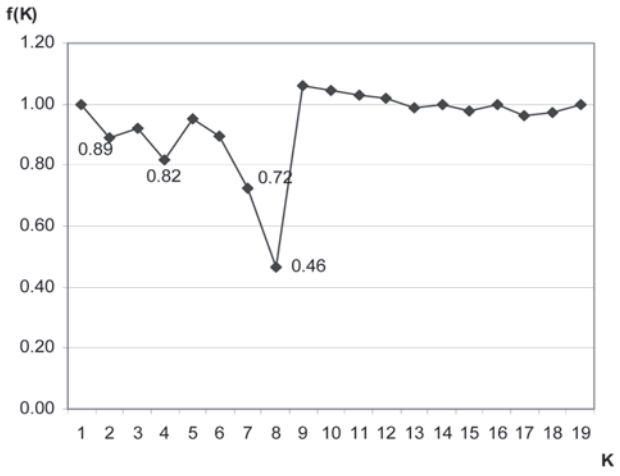
(j)





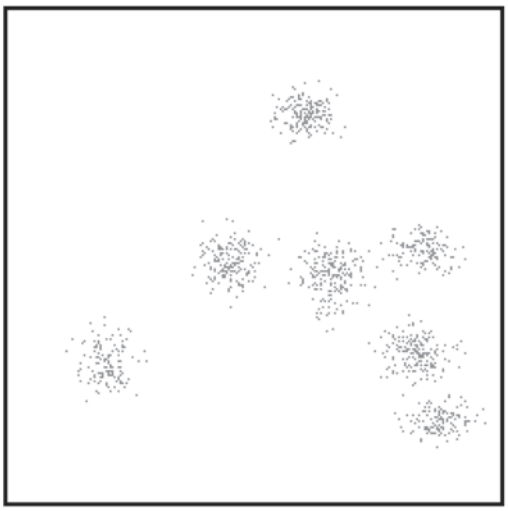
(k)

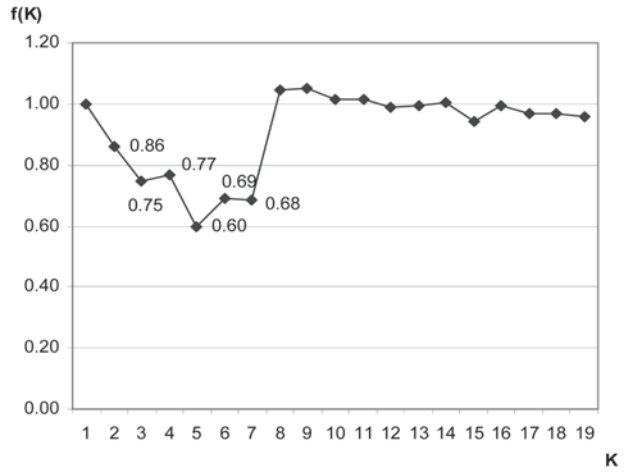




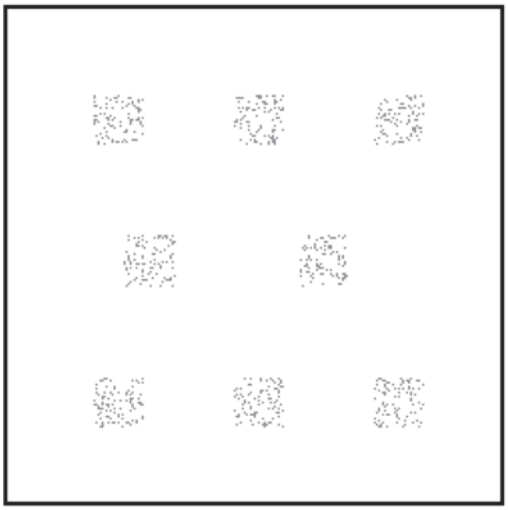
(l)

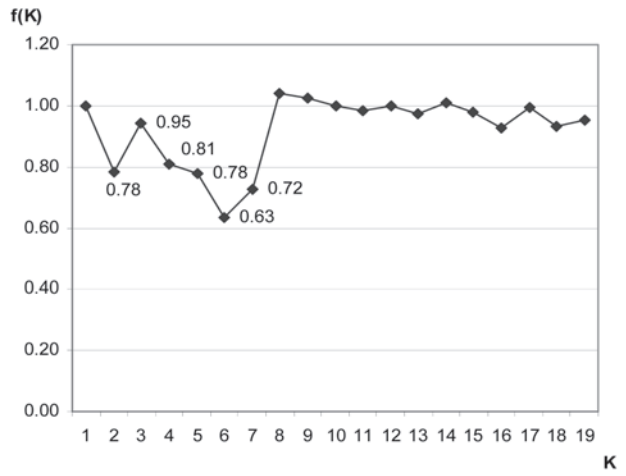
Fig. 6 Continued



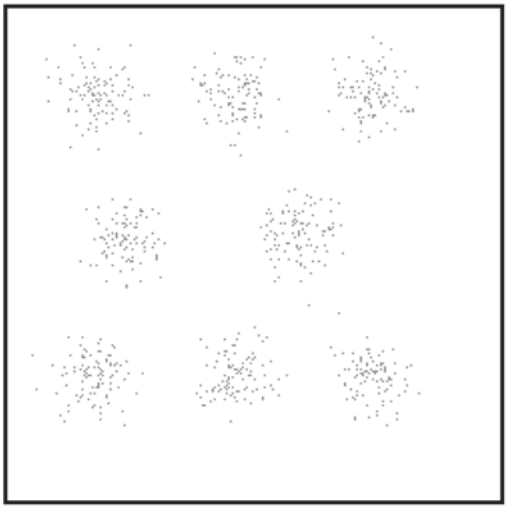


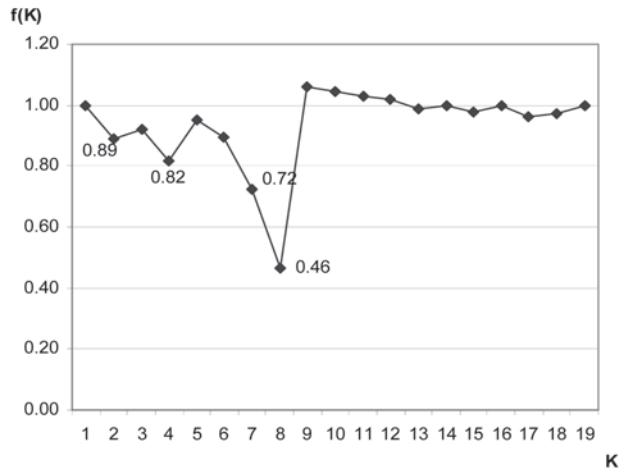
(m)





(n)





(o)

Fig. 6 Continued

from the average value for any strong recommendations to be made to the user. By comparing the values of $f(K)$ in Figs 6a and c, it can be seen that α_K reduces the effect of the data set dimensions on the evaluation function.

For the data set in Fig. 6d, again, all objects are concentrated in a single region with a normal distribution. The $f(K)$ plot for this data set suggests correctly that, when $K = 1$, the clustering result is the most suitable for this data set.

The data sets in Figs 6e and f are created by two generators that have normal distributions. In Fig. 6e, the two generators have an overlapping region but, in Fig. 6f, they are well separated. Note

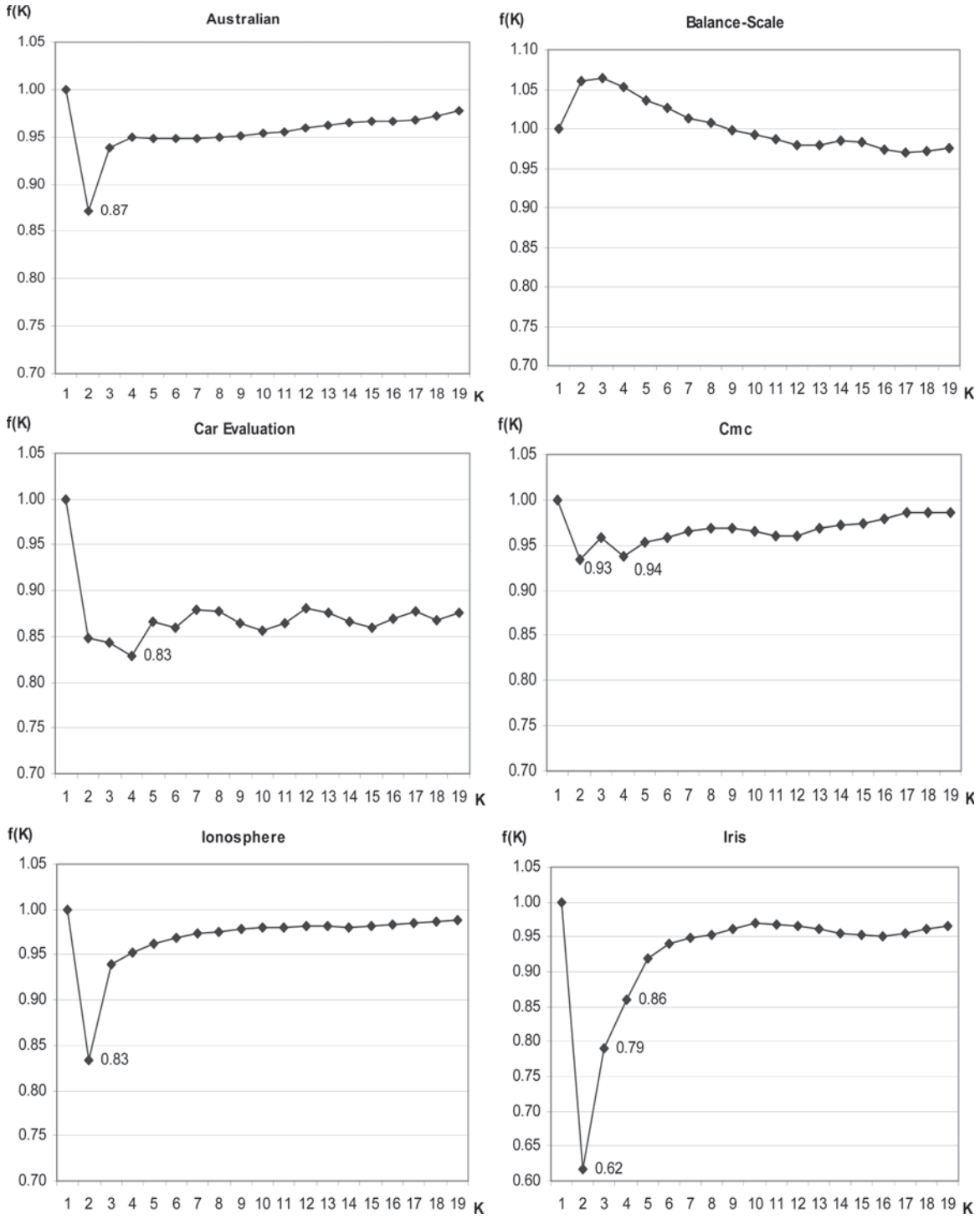


Fig. 7 $f(K)$ for the 12 benchmark data sets

that the value for $f(2)$ in the latter figure is much smaller than in the former.

The data sets in Figs 6g and h have three recognizable regions. From the corresponding graphs, $f(K)$ suggests correct values of K for clustering these data sets.

Three different generators that create object groupings with a normal distribution are used to form the data set in Fig. 6i. In this case, $f(K)$ suggests the value 2 or 3 for K. Because two of these three generators create object groupings that overlap, $f(2)$ is smaller than $f(3)$. This means that the data

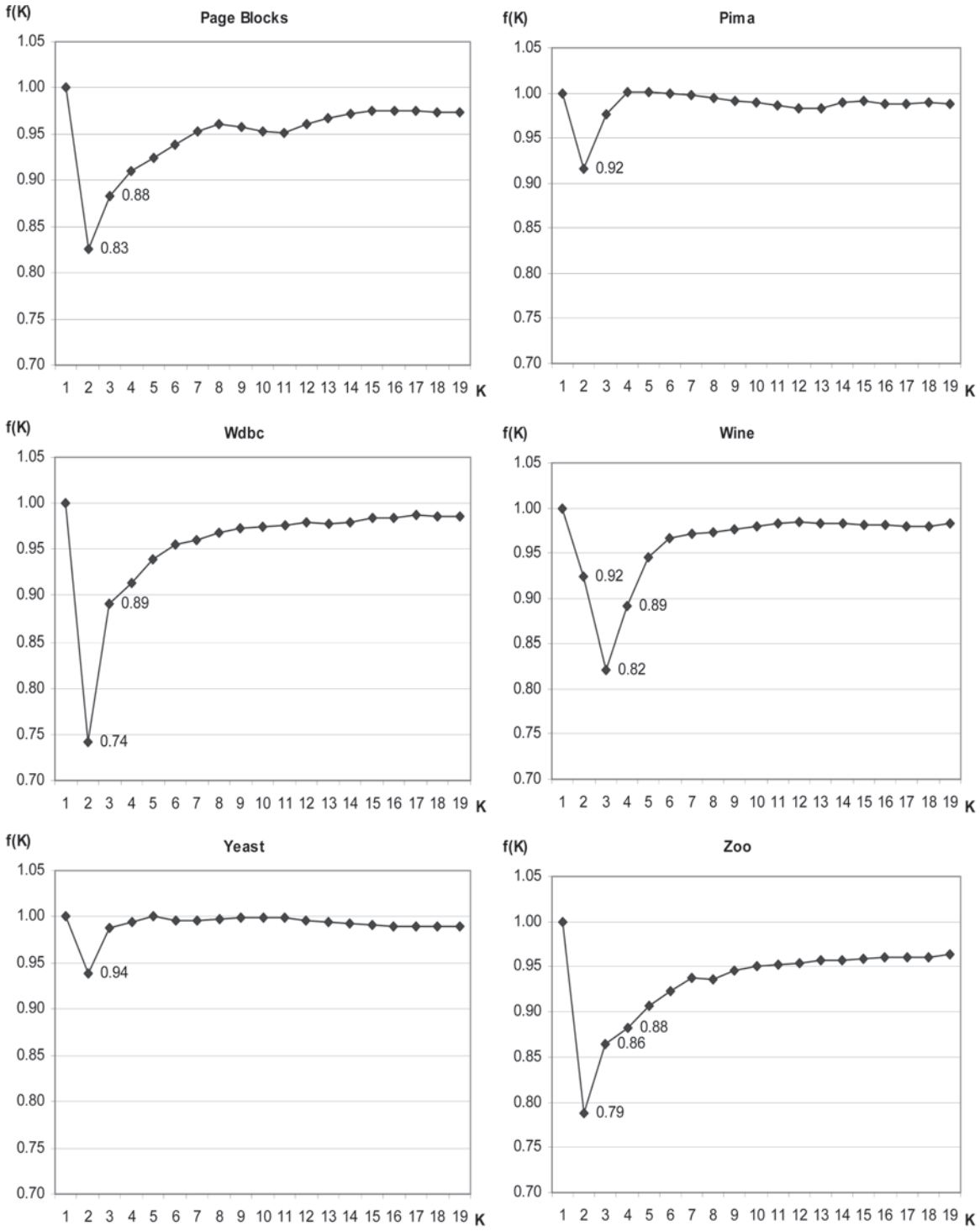


Fig. 7 Continued

have only two clearly defined regions, but $K = 3$ could also be used to cluster the objects.

Figures 6j and k illustrate how the level of detail could affect the selection of K. $f(K)$ reaches minimum values at $K = 2$ and 4 respectively. In such cases, users could select the most appropriate value of K based on their specific requirements. A more complex case is shown in Fig. 6l where there is a possible K value of 4 or 8. The selection of a particular K will depend on the requirements of the specific application for which the clustering is carried out.

The data sets in Figs 6m–o have well-defined regions in the object space, each of which has a different distribution, location, and number of objects. If the minimum value of $f(K)$ is used to cluster the objects, K will be different from the number of generators utilized to create them (as in the case of the clusters in Fig. 6o or the number of object groupings that could be identified visually (as in the case of the clusters in Figs 6m and n). The reason for the difference varies with different cases. For example, it could be considered that there are five clusters in Fig. 6m because the cluster distances are smaller for the two leftmost pairs of clusters than for others and the clusters in those pairs could be merged together. However, no simple explanation could be given for the cases shown in Figs 6n and o. This highlights the fact that $f(K)$ should only be used to suggest a guide value for the number of clusters and the final decision as to which value to adopt has to be left at the discretion of the user.

From the graphs in Fig. 6, a conclusion could be made that any K with corresponding $f(K) < 0.85$ could be recommended for clustering. If there is not a value with corresponding $f(K) < 0.85$, $K = 1$ is selected.

The proposed function $f(K)$ is also applied to 12 benchmarking data sets from the UCI Repository *Machine Learning Databases* [**31**]. Figure 7 shows how the value of $f(K)$ varies with K. If a threshold of 0.85 is selected for $f(K)$ (from the study on the artificial data sets), the numbers of clusters recommended for each of these data sets are given as in Table 2. $K = 1$ means that the data distribution is very close to the standard uniform distribution. The values recommended using $f(K)$ are very small because of the high correlation between the attributes of these data sets, very similar to that shown in Fig. 6e. This can be verified by examining two attributes at a time and plotting the data sets in two dimensions.

The above experimental study on 15 artificial and 12 benchmark data sets has demonstrated the robustness of $f(K)$. The evaluation function converges in most cases to 1 when K increases above 9.

Table 2 The recommended number of clusters based on $f(K)$

Data sets	Proposed number of clusters
Australian	1
Balance-scale	1
Car evaluation	2, 3, 4
Cmc	1
Ionosphere	2
Iris	2, 3
Page blocks	2
Pima	1
Wdbc	2
Wine	3
Yeast	1
Zoo	2

6 CONCLUSION

Existing methods of selecting the number of clusters for K-means clustering have a number of drawbacks. Also, current methods for assessing the clustering results do not provide much information on the performance of the clustering algorithm.

A new method to select the number of clusters for the K-means algorithm has been proposed in the paper. The new method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The proposed method can suggest multiple values of K to users for cases when different clustering results could be obtained with various required levels of detail. The method could be computationally expensive if used with large data sets because it requires several applications of the K-means algorithm before it can suggest a guide value for K. The method has been validated on 15 artificial and 12 benchmark data sets. Further research is required to verify the capability of this method when applied to data sets with more complex object distributions.

ACKNOWLEDGEMENTS

This work was carried out as part of the Cardiff Innovative Manufacturing Research Centre Project supported by the Engineering and Physical Sciences Research Council and the SUPERMAN Project supported by the European Commission and the Welsh Assembly Government under the European Regional Development Fund programme. The authors are members of the I*PROMS Network of Excellence funded by the European Commission.

REFERENCES

1 **Han, J.** and **Kamber, M.** *Data Mining: Concepts and Techniques*, 2000 (Morgan Kaufmann, San Francisco, California).

2 **Al-Daoud, M. B., Venkateswarlu, N. B.,** and **Roberts, S. A.** Fast *K*-means clustering algorithms. Report 95.18, School of Computer Studies, University of Leeds, June 1995.

3 **Al-Daoud, M. B., Venkateswarlu, N. B.,** and **Roberts, S. A.** New methods for the initialisation of clusters. *Pattern Recognition Lett.*, 1996, **17**, 451–455.

4 **Alsabti, K., Ranka, S.,** and **Singh, V.** An efficient *K*-means clustering algorithm. In Proceedings of the First Workshop on *High-Performance Data Mining*, Orlando, Florida, 1998; ftp://ftp.cise.ufl.edu/pub/faculty/ranka/Proceedings.

5 **Bilmes, J., Vahdat, A., Hsu, W.,** and **Im, E. J.** Empirical observations of probabilistic heuristics for the clustering problem. Technical Report TR-97-018, International Computer Science Institute, Berkeley, California.

6 **Bottou, L.** and **Bengio, Y.** Convergence properties of the *K*-means algorithm. *Adv. Neural Infn Processing Systems*, 1995, **7**, 585–592.

7 **Bradley, S.** and **Fayyad, U. M.** Refining initial points for *K*-means clustering. In Proceedings of the Fifteenth International Conference on *Machine Learning* (*ICML '98*) (Ed. J. Shavlik), Madison, Wisconsin, 1998, pp. 91–99 (Morgan Kaufmann, San Francisco, California).

8 **Du, Q.** and **Wong, T-W.** Numerical studies of MacQueen's *K*-means algorithm for computing the centroidal Voronoi tessellations. *Int. J. Computers Math. Applics*, 2002, **44**, 511–523.

9 **Castro, V. E.** and **Yang, J.** A fast and robust general purpose clustering algorithm. In Proceedings of the Fourth European Workshop on *Principles of Knowledge Discovery in Databases and Data Mining (PKDD 00)*, Lyon, France, 2000, pp. 208–218.

10 **Castro, V. E.** Why so many clustering algorithms? *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2002, **4**(1), 65–75.

11 **Fritzke, B.** The LBG-U method for vector quantization – an improvement over LBG inspired from neural networks. *Neural Processing Lett.*, 1997, **5**(1), 35–45.

12 **Hamerly, G.** and **Elkan, C.** Alternatives to the *K*-means algorithm that find better clusterings. In Proceedings of the 11th International Conference on *Information and Knowledge Management* (*CIKM 02*), McLean, Virginia, 2002, pp. 600–607.

13 **Hansen, L. K.** and **Larsen, J.** Unsupervised learning and generalisation. In Proceedings of the IEEE International Conference on *Neural Networks*, Washington, DC, June 1996, pp. 25–30 (IEEE, New York).

14 **Ishioka, T.** Extended *K*-means with an efficient estimation of the number of clusters. In Proceedings of the Second International Conference on *Intelligent Data Engineering and Automated Learning (IDEAL 2000)*, Hong Kong, PR China, December 2000, pp. 17–22.

15 **Kanungo, T., Mount, D. M., Netanyahu, N., Piatko, C., Silverman, R.,** and **Wu, A.** The efficient *K*-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Analysis Mach. Intell.* 2002, **24**(7), 881–892.

16 **Pelleg, D.** and **Moore, A.** Accelerating exact *K*-means algorithms with geometric reasoning. In Proceedings of the Conference on *Knowledge Discovery in Databases (KDD 99)*, San Diego, California, 1999, pp. 277–281.

17 **Pelleg, D.** and **Moore, A.** *X*-means: extending *K*-means with efficient estimation of the number of clusters. In Proceedings of the 17th International Conference on *Machine Learning (ICML 2000)*, Stanford, California, 2000, 727–734.

18 **Pena, J. M., Lazano, J. A.,** and **Larranaga, P.** An empirical comparison of four initialisation methods for the *K*-means algorithm. *Pattern Recognition Lett.*, 1999, **20**, 1027–1040.

19 *SPSS Clementine Data Mining System. User Guide Version 5*, 1998 (Integral Solutions Limited, Basingstoke, Hampshire).

20 *DataEngine 3.0 – Intelligent Data Analysis – an Easy Job*, Management Intelligenter Technologien GmbH, Germany, 1998; http://www.mitgmbh.de.

21 **Kerr, A., Hall, H. K.,** and **Kozub, S.** *Doing Statistics with SPSS*, 2002 (Sage, London).

22 *S-PLUS 6 for Windows Guide to Statistics*, Vol. 2, Insightful Corporation, Seattle, Washington, 2001; http://www.insightful.com/DocumentsLive/23/44/statman2.pdf.

23 **Hardy, A.** On the number of clusters. *Comput. Statist. Data Analysis*, 1996, **23**, 83–96.

24 **Theodoridis, S.** and **Koutroubas, K.** *Pattern Recognition*, 1998 (Academic Press, London).

25 **Halkidi, M., Batistakis, Y.,** and **Vazirgiannis, M.** Cluster validity methods. Part I. *SIGMOD Record*, 2002, **31**(2); available online http://www.acm.org/sigmod/record/.

26 **Kothari, R.** and **Pitts, D.** On finding the number of clusters. *Pattern Recognition Lett.*, 1999, **20**, 405–416.

27 **Cai, Z.** Technical aspects of data mining. PhD thesis, Cardiff University, Cardiff, 2001.

28 **Lindeberg, T.** *Scale-space Theory in Computer Vision*, 1994 (Kluwer Academic, Boston, Massachusetts).

29 **Pham, D. T., Dimov, S. S.,** and **Nguyen, C. D.** Incremental *K*-means algorithm. *Proc. Instn Mech. Engrs, Part C: J. Mechanical Engineering Science*, 2003, **218**, 783–795.

30 **Tibshirani, R., Walther, G.,** and **Hastie, T.** Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Department of Statistics, Stanford University, California, 2000.

31 **Blake, C., Keogh, E.,** and **Merz, C. J.** *UCI Repository of Machine Learning Databases, Irvine, California*. Department of Information and Computer Science, University of California, Irvine, California, 1998.

APPENDIX

Notation

A, B	clusters
$d(x_{jt}, w_j)$	distance between object x_{jt} and the centre w_j of cluster j
$f(K)$	evaluation function
G_A, G_B	generators
I_j	distortion of cluster j
K	number of clusters
N	number of objects in the data set
N_d	number of data set attributes (the dimension of the data set)
N_j	number of objects belonging to cluster j
P_{G_A}, P_{G_B}	probabilities that X is created by G_A or G_B respectively
P_{C_A}, P_{C_B}	probabilities that X is clustered into A or B respectively
S_K	sum of all distortions with K being the specified number of clusters
X	object
x_{jt}	object belonging to cluster j
w_j	centre of cluster j
α_K	weight factor