A New Cluster Validity Index Based on Fuzzy Granulation-degranulation Criterion

Sriparna Saha, *Student Member, IEEE* and Sanghamitra Bandyopadhyay, *Senior Member, IEEE*
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India-700108
Email: {sriparna_r, sanghami}@isical.ac.in

***Abstract*—Identification of correct number of clusters and the corresponding partitioning are two important considerations in clustering. In this paper, a new fuzzy quantization-dequantization criterion is used to propose a cluster validity index named Fuzzy Vector Quantization based validity index, *FVQ* index. This index identifies how well the formed cluster centers represent that particular data set. In general, most of the existing validity indices try to optimize the total variance of the partitioning which is a measure of compactness of the clusters so formed. Here a new kind of error function which reflects how well the formed cluster centers represent the whole data set is used as the goodness of the obtained partitioning. This error function is monotonically decreasing with increase in the number of clusters. Minimum separation between two cluster centers is used here to normalize the error function. The well-known genetic algorithm based K-means clustering algorithm (GAK-means) is used as the underlying partitioning technique. The number of clusters is varied from 2 to $\sqrt{N}$ where N is the total number of data points present in the data set and the values of the proposed validity index is noted down. The minimum value of the *FVQ* index over these $\sqrt{N}-1$ partitions corresponds to the appropriate partitioning and the number of partitions as indicated by the validity index. Results on five artificially generated and three real-life data sets show the effectiveness of the proposed validity index. For the purpose of comparison the cluster number identified by a well-known cluster validity index, XB-index, for the above mentioned eight data sets are also reported.**

***Index Terms*—Unsupervised classification, cluster validity index, fuzzy vector quantization**

I. Introduction

Clustering [1] is a core problem in data-mining with innumerable applications spanning many fields. The two fundamental questions that need to be addressed in any typical clustering scenario are: (i) how many clusters are actually present in the data, and (ii) how real or good the clustering itself. That is, whatever may be the clustering technique, one has to determine the number of clusters and also the validity of the clusters formed [2]. The measure of validity of clusters should be such that it will be able to impose an ordering of the clusters in terms of its goodness. In other words, if $U_1, U_2, \ldots, U_m$ be the m partitions of X, and the corresponding values of a validity measure be $V_1, V_2, \ldots V_m$, then $V_{k1} \geq V_{k2} \geq \ldots V_{km}, \forall ki \in 1, 2, \ldots, m,\ i = 1, 2, \ldots, m$ will indicate that $U_{k1} \uparrow \ldots \uparrow U_{km}$. Here '$U_i \uparrow U_j$' indicates that partition U_i is a better clustering than U_j. Note that a validity measure may also define a decreasing sequence instead of an increasing sequence of $V_{k1}, \ldots, V_{km}$. Several cluster validity indices have been proposed in the literature. These are Davies-Bouldin (DB) index [3], Dunn's index [4], Xie-Beni (XB) index [5], I-index [6], CS-index [7], etc., to name just a few. Some of these indices have been found to be able to detect the correct partitioning for a given number of clusters, while some can determine the appropriate number of clusters as well. Milligan and Cooper [8] have provided a comparison of several validity indices for data sets containing distinct non-overlapping clusters while using only hierarchical clustering algorithms. Maulik and Bandyopadhyay [6] evaluated the performance of four validity indices, namely, the Davies-Bouldin index [3], Dunn's index [4], Calinski-Harabasz index [6], and a new index $\mathcal{I}$, in conjunction with three different algorithms viz. the well-known K-means [1], single-linkage algorithm [1] and a SA-based clustering method [6].

Cluster properties such as compactness (or variation) and separation (or isolation) are often considered as major characteristics by which to validate clusters. Compactness is an indicator of the variation or the scattering of the data within a particular cluster, and separation is an indicator of the isolation of clusters from one another. In this paper a new approach has been adopted to validate the clusters formed. Here a fuzzy granulation and degranulation criterion proposed recently in [9] is used to validate the partitioning obtained. In the fuzzy granulation-degranulation criterion [9], the vectors in the code book are used to encode the original data in terms of the membership values. During decoding, a given vector is expressed as a function of the membership values and the cluster centers. The dissimilarity between the actual vector and the approximated vector at the decoder side is called the quantization error. In this paper the idea of fuzzy quantization-dequantization is used to measure how well the cluster centers, formed by a particular clustering algorithm for a data, represent the whole data set. Here these cluster centers are regarded as the representatives of the entire data set. Next, the final membership values of the data points present in the data set with respect to these cluster centers are calculated. Now based on these membership values and the cluster centers, each data point is approximated. The Euclidean distance between the approximated point and the original point is the error for that particular point. The average error (V) of the entire data set represents how well the partitioning is. It is easy to understand that with the increase in the number of clusters this quantization error decreases. Thus a normalization is done in order to get rid of the monotonically decreasing property of this error function thereby yielding a fuzzy vector quantization based validity index, referred to as *FVQ* index.

The well-known genetic algorithm based K-means clustering technique (GAK-means clustering technique) [10] is used as the underlying clustering algorithm. The number of clusters is varied from K_{min} to K_{max}. As a result, total $(K_{max} - K_{min} + 1)$ partitions will be generated, $U^*_{K_{min}}, U^*_{K_{min}+1} \ldots U^*_{K_{max}}$, with the corresponding validity index values computed as $V_{K_{min}}, V_{K_{min}+1} \ldots V_{K_{max}}$. Let $K^* = \text{a}rgopt_{i=K_{min} \ldots K_{max}}[V_i]$. Therefore, according to index V, K^* is the correct number of clusters present in the data. The corresponding U^*_K may be obtained by using a suitable clustering technique with the number of clusters set to K^*. The tuple $< U^*_{K^*}, K^* >$ is presented as the solution to the clustering problem. The effectiveness of the newly proposed cluster validity index, *FVQ* index, compared to a well-known cluster validity index, XB-index [5] is shown in identifying number of clusters from five artificially generated and three real-life data sets of varying complexities.

II. The Fuzzy Vector Quantization Method

In fuzzy vector quantization [9], the code book is formed of some vectors $\{v_1, v_2, \ldots v_K\}$. These are obtained by optimizing an error function after application of some optimization techniques. In general, code book consists of the elements of the data which approximate the whole data set appropriately. In [9], particle swarm optimization is used as the underlying optimization technique. The following discussion is based on the fuzzy vector quantization-dequantization approach proposed in [9].

A way of encoding a particular data point $\overline{x}$ in the data set can be represented by the collection of membership values to the different clusters. We require that the corresponding membership degrees $u_i(\overline{x}), i = 1, 2, \ldots K$ are confined to the unit interval and sum up to 1. The membership values are calculated by minimizing the following performance index

$$Q_1(x) = \sum_{i=1}^{K} u_i^m(x) \|\overline{x} - \overline{v}_i\|^2 \tag{1}$$

subject to the following constraints already stated above, that is

$$u_i(\overline{x}) \in [0, 1], \quad \sum_{i=1}^{K} u_i(\overline{x}) = 1 \tag{2}$$

The distance function is denoted by $\| \|^2$. The fuzzification coefficient $(m, m > 1)$, shown in the above expression is used to adjust the level of contribution of the prototypes to the result of representation. The collection of K weights $\{u_i(\overline{x})\},\ i = 1, \ldots K$ along with the cluster centers are used to represent a particular data point $\overline{x}$.

The minimization of Equation 1 is straightforward and follows a standard way of transforming the problem to unconstrained optimization using Lagrange multipliers. After rewriting the Equation 1 by accommodating the constraint in the form of the Lagrangian multiplier (λ), we obtain

$$Q_1(x) = \sum_{i=1}^{K} u_i^m(x) \|\overline{x} - \overline{v}_i\|^2 - \lambda(\sum_{i=1}^{K} u_i(\overline{x}) - 1) \tag{3}$$

The resulting system of equations leading to the minimum of Q comes in the form

$$\frac{dQ}{d\lambda} = 0, \qquad \frac{dQ}{du_i(\overline{x})} = 0 \tag{4}$$

After solving the equations with respect to λ and $u_i(\overline{x})$, the resulting weights (membership degrees) become

$$u_i(\overline{x}) = \frac{1}{\sum_{i=1}^{K}(\|\overline{x} - \overline{v}_i\| / \|\overline{x} - \overline{v}_j\|)^{2/(m-1)}} \tag{5}$$

where $i = 1, 2, \ldots K$. Here, the fuzzification coefficient, m is chosen equal to 2, though the importance of its proper choice is studied in [9].

Thus each data point is represented by the K membership values $u_i(\overline{x}),\ i = 1, \ldots K$ computed by Equation 5 and with the help of K cluster centers.

Now, these computed membership values and the cluster prototypes are used to approximate each data point $\overline{x}$. Approximation is based on some aggregation of the cluster centers and the associated membership grades $u_i(\overline{x})$. The way of forming $\overline{x'}$ is accomplished through a minimization of the following expression.

$$Q_2(\overline{x'}) = \sum_{i=1}^{K} u_i^m \|\overline{x'} - \overline{v}_i\|^2 \tag{6}$$

If the Euclidean distance is used to measure the distance between the prototypes and $\overline{x'}$, the problem of unconstrained optimization leads to a straightforward solution expressed as a convex combination of the prototypes

$$\overline{x'} = \frac{\sum_{i=1}^{K} u_i^m \overline{v}_i}{\sum_{i=1}^{K} u_i^m} \tag{7}$$

where the corresponding prototypes are weighted by the membership degrees. Then the total error due to clustering is calculated as follows.

$$\mathit{quan\_error} = \sum_{i=1}^{N} \|\overline{x}_i - \overline{x'}_i\|^2 \tag{8}$$

where N is the total number of points in the data set. It is shown in [9] that the quality of reconstruction depends on a number of essential parameters of the scheme including the size of the codebook (i.e., here the number of cluster centers) as well as the value of the fuzzification coefficient (m).

III. Proposed Cluster Validity Index

In case of clustering, the codebook consists of the cluster centers formed by a particular clustering technique. The membership values of the data points to different clusters are calculated using these cluster centers. Then, each data point is supposed to be well-represented by using these membership values and the cluster centers formed thus far.

If the obtained cluster centers are able to present global view of the data set, then the quantization error would be small. This error function is used here to validate a particular partitioning. But due to monotonically decreasing nature of the error function, here another factor is also taken into consideration, i.e., the minimal separation between any two cluster centers.

The proposed cluster validity index is defined below: Let the cluster centers of K clusters are represented by $\overline{v}_i$, $i = 1, \dots K$. Let C_i, $i = 1, \dots K$ represents the set of points which are in ith cluster and N represents total number of points in the data set. Then *FVQ* index is defined as follows.

$$FVQ = \frac{quan\_error}{N(\min_{i,k=1,\dots K, i \neq k} d_e(\overline{v}_i, \overline{v}_k))} \tag{9}$$

Here *quan_error* is computed by Equation 8. The total *quan_error* is divided by the total number of points, N in order to obtain the average quantization error. $d_e(\overline{v}_i, \overline{v}_k)$ represents the Euclidean distance between two cluster centers $\overline{v}_i$, and $\overline{v}_k$, i.e., the denominator of the obtained cluster center represents the minimum distance between any two cluster centers. The most desirable partition (or, an optimal value of K) is obtained by minimizing *FVQ* index over $K = 2, 3, \dots K_{max}$.

A. Explanation

As formulated in Equation 9, *FVQ* index is a composition of two factors. These are the average *quan_error* (the term $\frac{1}{N}$ is used to find the average quantization error over N number of data points) and $D_K = \min_{i,k=1,\dots K, i \neq k} d_e(\overline{z}_i, \overline{z}_k)$. The first factor denotes the average quantization error obtained due to representation of the data points using the cluster centers and the membership values. A minimum of this value indicates that the formed cluster centers represent the whole data set properly. This decreases as the number of cluster, K increases. But as *FVQ* needs to be minimized for obtaining proper number of clusters, so it will prefer to increase the value of K. Finally the second factor, D_K, measuring the minimum separation between a pair of clusters, decreases with increase in the value of K. In an ideal partitioning, the cluster centers should be well-separated. As the proposed *FVQ* index needs to be minimized, so it will prefer to increase the value of D_K making the minimum distance between any two cluster centers high. Thus as the two factors of the proposed *FVQ* index are complementary in nature, so they are expected to compete and balance each other critically for determining the proper partitioning.

IV. GAK-means: Clustering Algorithm Used for Segmentation

GAK-means [10] clustering algorithm is developed in order to get rid of the limitations of the well-known K-means algorithm to get struck at suboptimal solutions. Here the searching capability of GAs has been used for the purpose of appropriately determining a fixed number K of cluster centers in R^N; thereby suitably clustering the set of N unlabeled points. Each string in the population of GA is a sequence of real numbers representing the K cluster centers. These cluster centers are initialized to K randomly chosen points from the data set. The fitness computation process consists of two phases. In the first phase, the clusters are formed according to the centers encoded in the chromosome under consideration. This is done by assigning each point x_i, $i = 1, 2, \dots N$, to one of the clusters C_j with center v_j such that $\|x_i - v_j\| < \|x_i - v_p\|$, $p = 1, 2, \dots K$, $p \neq j$. After the assignments are done, the cluster centers encoded in the chromosome are replaced by the mean points of the respective clusters. The clustering metric, M is calculated as $M = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - v_i\|$. The fitness of the chromosome (fit) is calculated as: $fit = \frac{1}{M}$. The objective of the GA is to maximize this fitness function in order to minimize the total variance of the partitioning. Roulette wheel selection is used to implement the proportional selection strategy. Single point crossover with a fixed crossover probability is used. Each chromosome undergoes mutation with a fixed probability. Each gene position in a chromosome is mutated with a value lying near to it. Elitism has been incorporated in GA. The processes of fitness computation, selection, crossover, and mutation are executed for a maximum number of generations. The best string seen up to the last generation provides the solution to the clustering problem.

V. Experimental Results

Several artificially generated and real-life data sets were used to experimentally demonstrate that the *FVQ* index is able to find the proper cluster number for different types of data sets. Here results have been shown only for three real-life and five artificially generated data sets. In this section at first the description of the data sets used for the experiment are provided in brief. Finally the results are discussed in detail.

A. Data Sets Used

The data sets that are used for the experiment are divided into 2 different groups.

1) Group 1: Consists of five data sets. These data sets are used in [11].
   a) *Data1*: This data set consists of 250 data points distributed over 5 spherically shaped clusters in 2-dimensional space. The clusters present here are highly overlapping, consisting of 50 data points each. This data set is shown in Figure 1.
   b) *Data2*: This data set consists of 400 data points in 3-dimensional space distributed over 4 hyperspherical disjoint clusters. Each cluster contains 100 data points. This data set is shown in Figure 2(a).
   c) *Data3*: This data set consists of 76 data points distributed over 3 clusters. This data set is shown in Figure 2(b).
   d) *Data4*: This data set consists of 500 data points distributed over 10 different clusters. Some clusters are overlapping in nature. Each cluster con-

sists of 50 data points. This data set is shown in Figure 3.

   e) *Data5*: This data set consists of 300 data points distributed over 6 clusters in 2-dimensional space. The clusters are of the same sizes. This data set is shown in Figure 4.

2) Group 2: Consists of three real life data sets. These are *Iris*, *Cancer* and *Newthyroid* data sets.

   a) *Iris*: Iris data set consists of 150 data points distributed over 3 clusters. Each cluster consists of 50 points. This data set represents different categories of irises characterized by four feature values [12]. It has three classes Setosa, Versicolor and Virginica.

   b) *Breast Cancer*: Here we use the Wisconsin Breast Cancer data set consisting of 683 sample points. Each pattern has nine features corresponding to clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. There are two categories in the data: malignant and benign. The two classes are known to be linearly separable.

   c) *Newthyroid*: The original database from where it has been collected is titled as Thyroid gland data ('normal', 'hypo' and 'hyper' functioning). Five laboratory tests are used to predict whether a patient's thyroid belongs to the class euthyroidism, hypothyroidism or hyperthyroidism. There are a total of 215 instances. Total number of attributes is five.

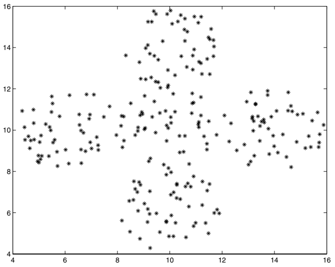

Fig. 1. *Data1*

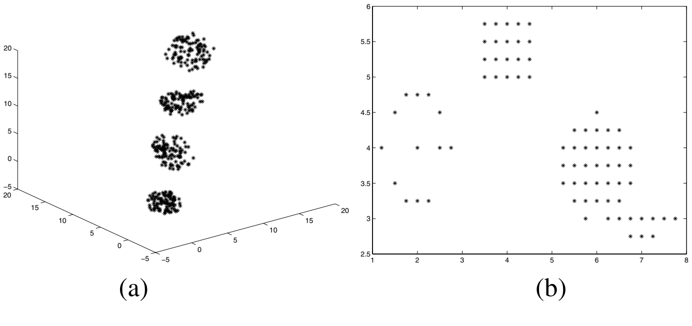

Fig. 2. (a) *Data2* (b) *Data3*

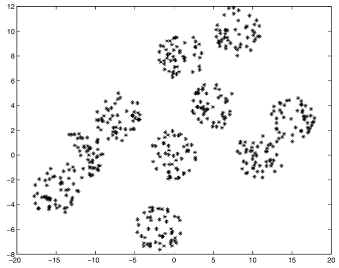

Fig. 3. *Data4*

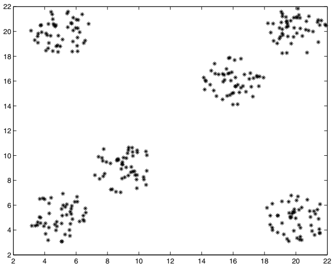

Fig. 4. *Data5*

B. Discussion of Results

The recently developed genetic algorithm based K-means clustering technique (GAK-means) is used as the underlying partitioning technique. The parameters of the genetic clustering algorithm (GAK-means) are as follows: population size is equal to 100 and maximum number of generations is equal to 30. The crossover and mutation probabilities are chosen as 0.8 and 0.01, respectively. The number of clusters (K) is varied from 2 to $\sqrt{N}$, where N is the total number of data points present in the data set and the variation of the *FVQ* index is noted down. Its minimum value indicates the appropriate algorithm and the appropriate number of clusters. For the purpose of comparison, the performance of a well-known cluster validity index, XB-index [5], is also tested on these eight artificially generated and real-life data sets. Table I shows the optimum values of both the validity

TABLE I

OPTIMAL VALUES OF *FVQ* INDEX AND XB-INDEX IN THE RANGE OF $K = 2, \ldots \sqrt{N}$ FOR ALL THE EIGHT DATA SETS. HERE OC DENOTES THE CLUSTER NUMBER CORRESPONDING TO WHICH THE INDEX ATTAINS ITS OPTIMUM VALUE. AC DENOTES THE ACTUAL NUMBER OF CLUSTERS PRESENT IN THE DATA SET.

Data set	AC	FVQ index		XB-index	
		OC	Value	OC	Value
Data1	5	5	50.23	4	0.1388
Data2	4	4	0.149	4	0.052
Data3	3	3	0.0845	3	0.083
Data4	10	10	0.189449	4	0.148
Data5	6	6	0.158	4	0.043
Iris	3	3	0.027863	2	0.066
Cancer	2	2	0.178842	2	0.15
Newthyroid	3	4	0.634777	4	0.1946

indices, *FVQ* index and XB-index, and the corresponding number of clusters obtained after application of the GAK-means clustering algorithm on different data sets.

Figures 5, 7, 6, 8 and 9 show, respectively, the partitionings obtained by GAK-means algorithm after application on *Data1*, *Data2*, *Data3*, *Data4* and *Data5* with the number of clusters identified by the proposed cluster validity index, *FVQ* index. It can be easily seen from Table I that the proposed *FVQ* index is able to detect the proper partition number from all the artificial data sets. The identified partitions are also perfect as seen from Figures 5, 6, 7, 8 and 9, respectively. XB-index is able to detect the proper number of clusters only from *Data2* and *Data3* (refer to Table I). For the other three data sets, it is not able to identify the proper partitioning and the proper number of clusters.

For the real-life data sets, no visualization is possible as these are higher-dimensional data sets. For both *Iris* and *Cancer* data sets, the proposed index is able to detect the proper partition number. In order to measure the goodness of the partitioning, the *Minkowski Score* [13] is calculated after application of GAK-means algorithm. This is a measure of the quality of a solution given the true clustering. Let T be the "true" solution and S the solution we wish to measure. Denote by n_{11} the number of pairs of elements that are in the same cluster in both S and T. Denote by n_{01} the number of pairs that are in the same cluster only in S, and by n_{10} the number of pairs that are in the same cluster in T. *Minkowski Score* (MS) is then defined as:

$$MS(T,S) = \sqrt{\frac{n_{01} + n_{10}}{n_{11} + n_{10}}}. \tag{10}$$

For MS, the optimum score is 0, with lower scores being "better". For *Iris* and *Cancer* data sets, MS scores of the partitionings corresponding to the optimum value of *FVQ* index are 0.602 ± 0.003 and 0.367 ± 0.003, respectively. For *Newthyroid* data set, the proposed *FVQ* index is unable to detect the proper partition number. It wrongly detects 4 as the proper cluster number. The MS score corresponding to this partitioning is 0.62 ± 0.02. The well-known XB-index is able to detect the proper cluster number only for *Cancer* data set (refer to Table I). For *Iris* data set, it identifies $K = 2$ as the proper number of clusters, which is also often obtained for many other methods for *Iris*. Table I shows that XB-index is also unable to identify the proper number of clusters for *Newthyroid* data set.

Figures 10, 11, and 12 show, respectively, the variations of the proposed *FVQ* index over the number of clusters for *Data1*, *Data3*, and *Iris* data sets, for the purpose of illustration.

VI. Discussion and Conclusion

In this paper a new cluster validity index, named *FVQ* index, is proposed which uses a new error function to validate the obtained partitions. Thus this index is capable of detecting both the proper number of clusters as well as the proper partitioning from a data set. The numerator of this validity index is based on the fuzzy vector quantization-dequantization criterion. This error function gives a quantitative measurement of how well the obtained cluster centers represent the whole data set. As like the fuzzy vector quantization, all the data points are represented by using the obtained cluster centers and their membership values. Then the Euclidean distance between the original data point and the approximated point provides the total approximation error of that particular data point due to clustering. The total average approximation error intuitively gives an idea how well the obtained cluster centers represent the whole data set. As this error function is monotonically decreasing with increase in the number of clusters, minimum separation between any two cluster centers is used to normalize this error function. Thus minimum value of the proposed validity index *FVQ* which is the ratio of the average quantization error and the minimum separation, corresponds to proper partitioning and the proper partition number. The effectiveness of the proposed index as compared to the well-known XB-index is shown in detecting proper partitioning from five artificially generated and three real-life data sets along with GAK-means clustering algorithm.

Future work includes use of some other distances in place of the Euclidean distance while calculating the membership values of different points to different clusters, so that the proposed index is able to detect some non convex/ convex symmetrical clusters other than hyperspherical ones.

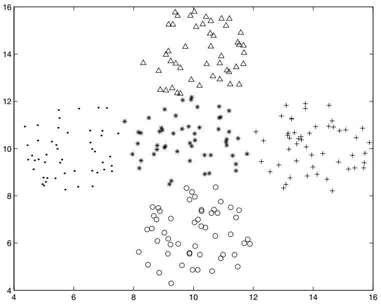

Fig. 5. Clustered *Data1* after application of GAK-means for $K = 5$

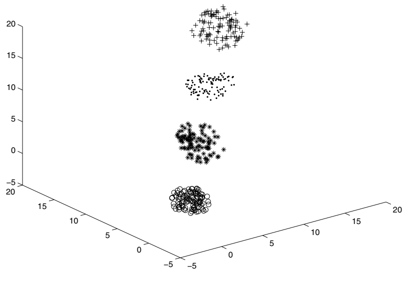

Fig. 6. Clustered *Data2* after application of GAK-means for $K = 4$

References

[1] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. London: Arnold, 2001.

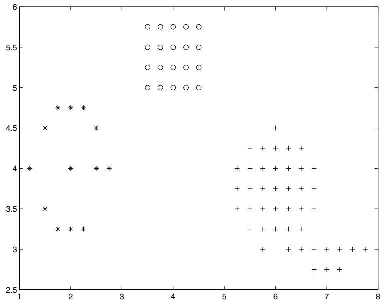

Fig. 7. Clustered *Data3* after application of GAK-means for $K = 3$

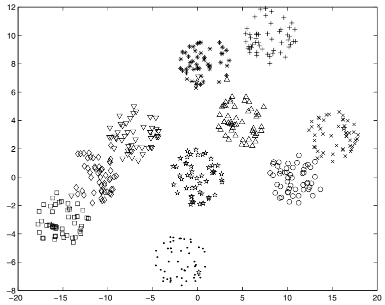

Fig. 8. Clustered *Data4* after application of GAK-means for $K = 10$

[2] R. C. Dubes and A. K. Jain, "Clustering techniques : The user's dilemma," *Pattern Recognition*, vol. 8, pp. 247–260, 1976.

[3] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions Patt. Anal. Mach. Intell.*, vol. 1, pp. 224–227, 1979.

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

[5] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 841–847, 1991.

[6] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[7] C. H. Chou, M. C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis and Applications*, vol. 7, pp. 205–220, 2004.

[8] G. W. Milligan and C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[9] W. Pedrycz and K. Hirota, "Fuzzy vector qunatization with the particle swarm optimization: A study in fuzzy granulation-degranulation information processing," *Signal Processing*, vol. accepted, doi:10.1016/j.sigpro.2007.02.001, 2007.

[10] U. Maulik and S. Bandyopadhyay, "Genetic algorithm based clustering technique," *Pattern Recog.*, vol. 33, pp. 1455–1465, 2000.

[11] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of clusters and application to image classification," *Pattern Recognition*, no. 2, pp. 1197–1208, 2002.

[12] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 3, pp. 179–188, 1936.

[13] A. Ben-Hur and I. Guyon, *Detecting Stable Clusters using Principal Component Analysis in Methods in Molecular Biology*, M. Brownstein and A. Kohodursky, Eds. Humana press, 2003.

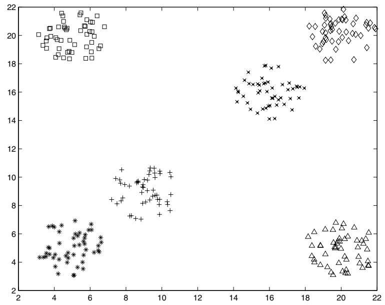

Fig. 9. Clustered *Data5* after application of GAK-means for $K = 6$

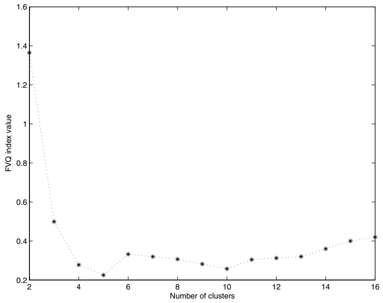

Fig. 10. Variation of the *FVQ* index value with number of clusters for *Data1*

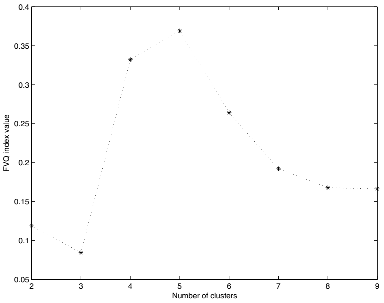

Fig. 11. Variation of the *FVQ* index value with number of clusters for *Data3*

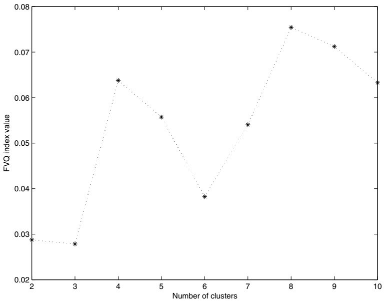

Fig. 12. Variation of the *FVQ* index value with number of clusters for *Iris*