Chapter 5

SIGT: Synthetic Image Generation Tool for Clustering Algorithms

A new automatic image generation tool is proposed in this chapter tailored specifically for the verification and comparison of different unsupervised image classification algorithms. The tool can be used to produce different images (in raw format) with different criteria based on user specification. The user specifies the number of clusters to be included in the image along with the probability distribution that governs a set of points that belong to different clusters. On the other hand, the tool can be used to verify the degree of approximation a new algorithm has been able to achieve compared to the original image. This allows for a scientific confident comparison between any new algorithm and existing algorithms. The usefulness of the tool is demonstrated in this chapter with reference to the well-known K-means clustering algorithm and the PSO-based clustering algorithm proposed in the previous chapter.

5.1 Need for Benchmarks

Researchers usually use their own data sets to test the performance of their clustering algorithms [Puzicha *et al*. 2000; Rosenberger and Chehdi 2000; Lorette *et al*. 2000; Boujemaa 2000; Huang 2002]. In addition, many researchers create their own synthetic data to test their algorithms. This approach makes the comparison between different clustering algorithms difficult. To address this problem, it is necessary to create a simple tool which will help researchers to create synthetic images. Researchers can then apply their clustering algorithm on these images and evaluate

the performance of these algorithms. Furthermore, researchers can agree to use some of these synthetic images as benchmarks making comparison between different clustering algorithms easier.

In this chapter, a new tool is proposed to generate benchmark images tailored specifically for clustering problems that have the capability to verify the clustering quality of any unsupervised image classification algorithm. This tool has the following benefits:

1. The tool can create synthetic images customized toward user-specific objectives. The user can create the images he/she wants with the desired complexity that suits his/her interests. The user specifies the number of clusters in advance to test the ability of the algorithm to find that number of clusters (especially in the case of unsupervised classification). Furthermore, the user can specify the degree of *closeness* (inter-cluster distances) between different clusters to control the complexity of different algorithms to be able to differentiate between close clusters. Finally, users are able, using the tool, to specify the distribution probability that govern the relationship between different points belonging to different clusters.
2. The tool can measure the quality of any clustering algorithm provided that it uses the tool's generated images and generate a thematic map image in a raw format. Hence, the user can measure the quality of a user-defined clustering algorithm to examine how efficient the algorithm is on different data sets.
3. According to the above benefits, the tool can be used to create a carefully crafted set of benchmark images. Hence, using SIGT, researchers can build common benchmark images to be used for comparison among different

clustering algorithms. The ability of the tool to easily create images with pre-defined clusters pushes towards this direction.

4. SIGT can be used to quantify the average performance of a user-specified clustering algorithm. This can be done by running the algorithm for a number of simulations on a library of benchmark images to statistically compare it to other algorithms.
5. This tool could be upgraded to generate a synthetic image from an existing image by relaxing some constraints. One way to do this is by calculating the histogram of the existing image and then composing a user defined separation period along the histogram, thus generating a modified cloned image.

Therefore, SIGT is best used as a preliminary test for any clustering algorithm (especially in the area of unsupervised image classification or segmentation). One of the advantages of the proposed tool is the ease with which one can measure the performance of a clustering algorithm, and compared its performance with other algorithms. The rest of the chapter is organized as follows: The proposed tool is described in detail in section 5.2. Section 5.3 provides experimental results verifying the applicability of the tool. Finally section 5.4 concludes the chapter.

5.2 SIGT: Synthetic Image Generation Tool

A synthetic image generation tool for clustering algorithms, SIGT, is proposed in this chapter to assist the unsupervised image classification research community's ability to compare and quantify the performance of different algorithms. The proposed tool

consists of two units: a synthetic image generator unit, and a clustering verification unit. This section provides a detailed description of each unit.

5.2.1 Synthetic Image Generator

With the synthetic image generation algorithm, the user can generate a synthetic image in raw format (converting an image from/to raw format can be achieved easily by using different graphic editing tools such as Adobe Photoshop) suitable for his/her objectives by specifying the following characteristics of the desired image:

- Size in pixels (i.e. the number of image pixels), N_p
- Dynamic range in bits (e.g. 8-bit image)
- Number of clusters, K
- Histogram characteristics:
  - Percentage of points in the image that belongs to each cluster, C'_k, $\forall\ k = 1,\ldots, K$.
  - Each cluster width in pixels (e.g. 10-pixels wide), w'_k, $\forall\ k = 1,\ldots, K$.
  - The probability distribution that govern points in each cluster, p_k, $\forall\ k = 1,\ldots, K$. The tool allows the user to select any of the following discrete random distributions [Leon-Garcia 1994; Devore 1995]: Uniform, Binomial, Geometric, and/or Poisson. These distributions represent the most common distributions. Therefore, the user can create an image with the histogram of his/her choice.

- The separation (in number of pixels) between adjacent clusters $S_{k,kk} \geq 0$ where k, kk = 1,…, K, i.e. the inter-cluster distance between two adjacent clusters $\boldsymbol{C}_k$ and $\boldsymbol{C}_{kk}$.

After specifying the above information, the synthetic image can be constructed according to the algorithm summarized in Figure 5.1.

For each cluster k = 1,…, K **do**

Distribute $C'_k . N_p/100$ pixels of cluster $\boldsymbol{C}_k$ according to p_k

Endloop

For each pixel z_p **do**

Assign z_p randomly to any cluster $\boldsymbol{C}_k$, where k = 1,…, K, based on the image histogram

Endloop

Create a thematic map image using

$$z_p = \frac{255k}{K} \qquad \forall\, p = 1,\ldots, N_p\,,\ k = 1,\ldots, K \tag{5.1}$$

where k is the index of cluster $\boldsymbol{C}_k$ to which z_p belongs

Figure 5.1: The synthetic image generator algorithm

The synthetic image generation algorithm works as follows: The first step constructs a synthetic image histogram by distributing N_p image pixels into the K clusters according to the cluster distribution specified by the user. The number of pixels in

cluster $\boldsymbol{C}_k$ can be determined by multiplying the percentage of points in the image belonging to cluster $\boldsymbol{C}_k$ by the number of image pixels N_p divided by 100 (i.e. $C'_k.N_p/100$). The second step assigns the synthetic image pixels randomly to the clusters according to the histogram created in the first step. Now, the synthetic image has been generated with the specified histogram. Finally, a thematic map image is generated according to equation (5.1). This map will help the user to verify the accuracy of any clustering algorithm when used in the clustering verification unit.

The synthetic image generator unit therefore generates two images: The first is the synthetic image representing the data set specified by the user, and the second is a thematic map of the generated synthetic image.

5.2.2 Clustering Verification Unit

The clustering verification unit verifies the average classification accuracy of the clustering algorithm being evaluated. This is done by comparing the thematic map generated from the synthetic image generation unit (or if the user has a thematic map representing the correct clustering) and the thematic map resulting from the clustering algorithm (this thematic map should be generated using equation (5.1)). This unit consists of two components, namely the clustering validity checker and the average performance analyzer. These components are described next.

5.2.2.1 Clustering Validity Checker Component

The clustering validity checker component verifies the average performance of any clustering algorithm with reference to a single image. It calculates the average

classification accuracy over a number of trials. The component requires the user to specify the following information:

- Dynamic range in bits.
- Original image name (e.g. synthetic image name generated from the first unit).
- Reference thematic map name representing the correct classification (e.g. thematic map generated from the first unit), TM_r.
- Executable program name of the clustering algorithm.
- Thematic map resulting from the clustering algorithm being examined (e.g. K-means, FCM, etc.), TM_c.
- Image size in pixels, N_p.
- Number of clusters, K.
- Number of runs of the clustering program, N_r.

The clustering verification algorithm is summarized in Figure 5.2.

For each pixel in the image, the cluster numbers as generated from TM_r and TM_c are compared. If the cluster numbers match, a counter, *mcount*, is incremented. Finally, the classification accuracy is calculated using equation (5.2). This algorithm is repeated N_r times followed by calculating the average classification accuracy.

The clustering verification unit produces a binary image showing the difference between the two thematic maps. This difference is represented as a white colored pixel for each incorrectly classified pixel. Furthermore, the unit calculates an accuracy percentage according to these differences.

mcount = 0 /* number of correctly classified pixels */

For each pixel z_p **do**

k = cluster number as in TM_r

kk = cluster number as in TM_C

If $k \neq kk$ **then**

Mark z_p as a white dot in the difference image

else

$mcount = mcount + 1$

Endif

Endloop

$$Classification\ accuracy = (mcount/N_p).100 \quad (5.2)$$

Figure 5.2: The clustering verification algorithm

5.2.2.2 Average Performance Analyzer Component

The average performance analyzer component verifies the average performance of any clustering algorithm with reference to a set of images specified by the user (i.e. a benchmark library). It calculates the average classification accuracy by running the program a user-specified number of trials on each image in the library. In this regards, the user should specify the following information:

- Executable program name of the clustering algorithm.
- Thematic map resulting from the clustering algorithm being examined (e.g. K-means, FCM, etc.), TM_c.
- Number of trials to run the clustering program, N_r.

Other information (e.g. the number of benchmarks, name of benchmark images, reference thematic maps, etc.) should be specified in an input file called *SIGT_classifier_input.txt*.

The clustering verification algorithm in section 5.2.2.1 is also used in this component. The only difference is that the clustering program is applied to each image in the benchmark library. Note that the tool deals only with raw format images. The next section presents results obtained from SIGT to illustrate its features.

5.3 Experimental Results

This section shows how SIGT can be used to standardize the unsupervised image classification verification process. For the sake of showing its applicability, two clustering algorithms were used, namely K-means and PSO-based clustering algorithms. The tool can generate 8-bit, flexible size images. A preliminary core of a benchmark library consisting of ten synthetic images with different levels of complexity and pixel intensity distribution were created. Table 5.1 shows user-specified parameters used in generating the images. The first and second images were generated such that they resemble very clear separated clusters. Therefore, it must be very easy for most clustering techniques to determine these clusters with great accuracy. The first image resembles a two-cluster image, while the second resembles a three-cluster image. The third synthetic image was generated to be slightly more difficult than the previous two in such a way that two of its three clusters are close to each other. The fourth image has three clusters adjacent to each other. On the other

hand, the fifth image has one more cluster than the fourth. In fact, the fourth and fifth images better approximate real-life images than the first three, which can only be considered for functionality verification rather than competitance of different clustering algorithms in early phases of algorithm creation. The remaining images are the most complex among all, where different difficulty levels are introduced in such a way that only competitive clustering algorithms will be able to efficiently cluster the regions in the images. A large number of adjacent clusters of different probability distributions were used in constructing these images.

For all the experiments, both K-means and *gbest* PSO-based clustering algorithms (using equation (4.6)) were averaged over 30 trials for each image in the benchmark library. The average classification accuracy and confidence interval (CI) are calculated (see Table 5.2). For the PSO-based clustering algorithm, 50 particles are used for 50 generations, $w_1 = w_2 = 0.3$ and $w_3 = 0.4$. The inertia weight, w, is set to 0.72, and $c_1 = c_2 = 1.49$. The velocities are clamped using $V_{max} = 255$.

Using the clustering verification unit of SIGT, the thematic maps obtained from both K-means and PSO clustering algorithms were compared with the thematic maps generated by the synthetic image generation unit. The images representing the difference in thematic maps are included in Table 5.2. The average classification accuracy, calculated using equation (5.2), and the confidence interval of both algorithms are included in Table 5.2 for each image.

It is observed that as the separation between adjacent clusters decreases, the classification accuracy becomes lower. Note that the PSO-based clustering algorithm performed better than K-means in all the cases except two (Image 8 and 10). The rationale for the poor performance of the PSO-based clustering algorithm when applied to Image 10 is the choice of w_1, w_2 and w_3. When the PSO-based clustering

algorithm was applied to Image 10 using $w_1 = w_2 = 0$ and $w_3 = 1$ (i.e. making the PSO fitness function similar to the objective function of K-means) the average classification accuracy significantly improved to reach 80.44% ± 7.411 with CI = [74.214, 86.674].

From the overall average performance, it can be concluded that the PSO-based clustering algorithm is in general better than the K-means clustering algorithm which verifies the results of chapter 4. To be able to make such a conclusion is one of the main benefits of SIGT.

Although the synthetic images represent no visually appealing shape, their histograms represent exactly what the user intends to test. The rationale behind this is that a clustering algorithm generally clusters pixels in a spectral domain (as represented in the histogram) rather than a spatial domain (as represented by the image shape). Therefore, the shape is generally not used, but rather the image histogram. However, in image segmentation the spatial information is important as already discussed in Section 3.2.

5.4 Conclusions

A new tool for synthetic image generation (SIGT) was proposed and implemented. The tool consists of two units: a synthetic image generator and a clustering verification unit. The first unit allows the user to create a synthetic image based on a user-specified histogram suitable for the required application. The second unit allows the user to measure the efficiency of a clustering algorithm. The main purpose of SIGT was to provide a simple and easy tool to generate synthetic images that can be used as benchmarks and to conduct a preliminary test on a clustering algorithm in

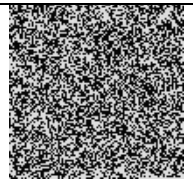
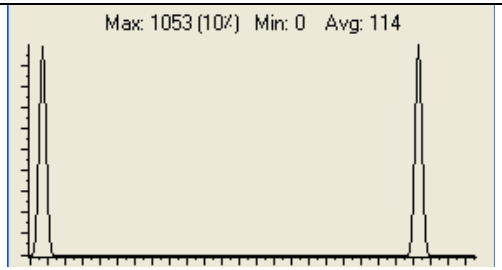
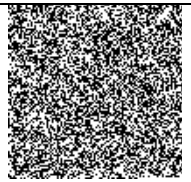
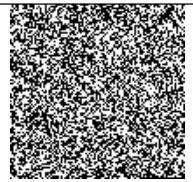
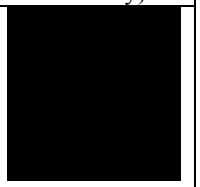
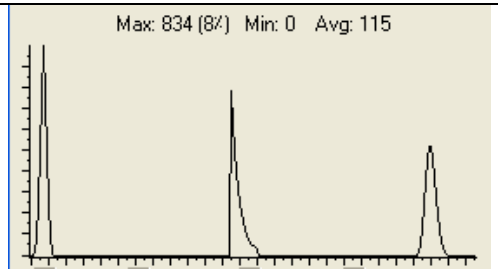
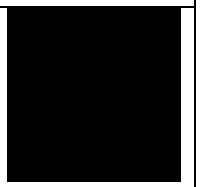
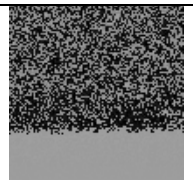
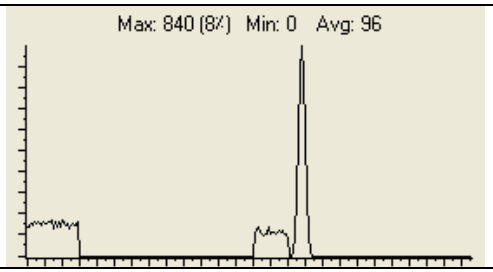
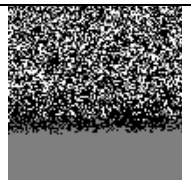
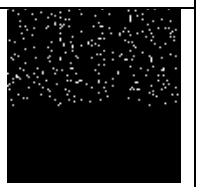
order to measure its performance and compare it with other clustering algorithms. Different features of SIGT were demonstrated by a set of experiments aided by a well-known clustering algorithm (K-means) and the recent PSO-based clustering algorithm. These experiments have shown that the tool can be used to generate a synthetic image based on a user-specified histogram, measure the quality of any clustering algorithm, and create benchmarks.

After showing the potential of PSO as a clustering algorithm and proposing a tool that can aid in evaluating the efficiency of a clustering algorithm, the next chapter will address the difficult problem of determining the "optimal" number of clusters in a data set.

Table 5.1: Synthetic image details and classification accuracy

Benchmark No.	*K*	***% of each cluster*** (C'_k)	***Cluster distribution*** (p_k)
1	2	1 (50%) 2 (50%)	1 (Binomial) 2 (Binomial)
2	3	1 (34%) 2 (33%) 3 (33%)	1 (Binomial) 2 (Geometric) 3 (Poisson)
3	3	1 (40%) 2 (20%) 3 (40%)	1 (Uniform) 2 (Uniform) 3 (Binomial)
4	3	1 (40%) 2 (20%) 3 (40%)	1 (Uniform) 2 (Uniform) 3 (Uniform)
5	4	1 (30%) 2 (20%) 3 (30%) 4 (20%)	1 (Uniform) 2 (Uniform) 3 (Uniform) 4 (Poisson)
6	10	1 (10%) 2 (5%) 3 (10%) 4 (10%) 5 (5%) 6 (10%) 7 (15%) 8 (10%) 9 (10 %) 10 (15%)	1 (Uniform) 2 (Uniform) 3 (Uniform) 4 (Poisson) 5 (Uniform) 6 (Binomial) 7 (Geometric) 8 (Uniform) 9 (Poisson) 10 (Binomial)
7	6	1 (20%) 2 (20%) 3 (15%) 4 (30%) 5 (5%) 6 (10%)	1 (Poisson) 2 (Binomial) 3 (Uniform) 4 (Uniform) 5 (Uniform) 6 (Uniform)
8	4	1 (25%) 2 (25%) 3 (25%) 4 (25%)	1 (Geometric) 2 (Binomial) 3 (Binomial) 4 (Uniform)
9	7	1 (20%) 2 (10%) 3 (35%) 4 (5%) 5 (15%) 6 (15%) 7 (10%)	1 (Uniform) 2 (Uniform) 3 (Uniform) 4 (Uniform) 5 (Uniform) 6 (Uniform) 7 (Binomial)
10	4	1 (25%) 2 (25%) 3 (25%) 4 (25%)	1 (Poisson) 2 (Binomial) 3 (Uniform) 4 (Uniform)

Table 5.2: Synthetic images, Histograms and Thematic Maps

Bench Mark No.	Synthetic Image	Histogram	K-Means Sample TM_c (Avg. Classification Accuracy±SD[a]) [CI]	PSO Sample TM_c (Avg. Classification Accuracy±SD) [CI]	Best K-Means TM Difference (Best Classification Accuracy)	Best PSO TM Difference (Best Classification Accuracy)
1			 100%±0 [100, 100]	 100%±0 [100, 100]	 (100%)	 (100%)
2			 100%±0 [100, 100]	 100%±0 [100, 100]	 (100%)	 (100%)
3			 79.58%±18.672 [72.901, 86.265]	 96.49%±0.491 [96.310, 96.662]	 (98.25%)	 (97.30%)

[a]SD stands for standard deviation

Table 5.2 (*continued*).

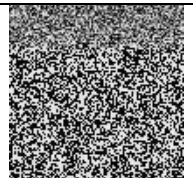
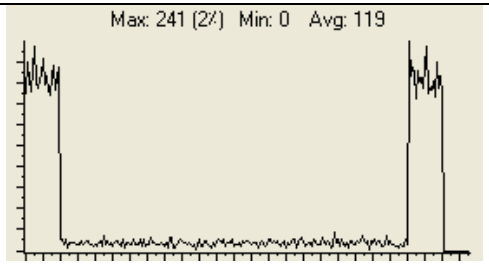
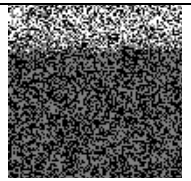
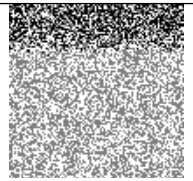
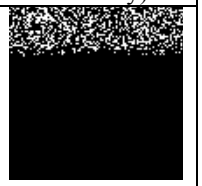
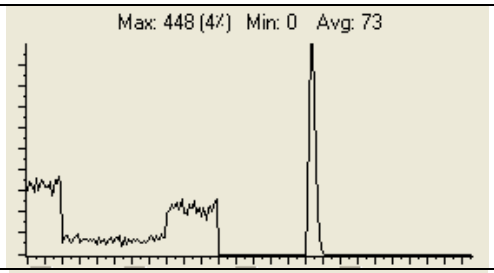
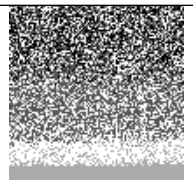
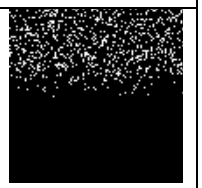
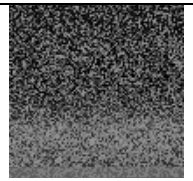
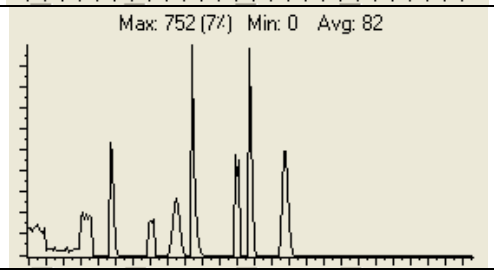
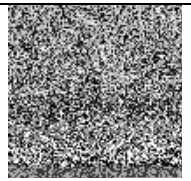
Bench Mark No.	Synthetic Image	Histogram	K-Means Sample TM_c (Avg. Classification Accuracy±SD) [CI]	PSO Sample TM_c (Avg. Classification Accuracy±SD) [CI]	Best K-Means TM Difference (Best Classification Accuracy)	Best PSO TM Difference (Best Classification Accuracy)
4			 90.56%±0 [90.560, 90.560]	 90.69%±0.060 [90.664, 90.707]	 (90.56%)	 (90.77%)
5			 91.75%±7.647 [89.018, 94.490]	 92.18%±0.318 [92.070, 92.298]	 (93.21%)	 (92.48%)
6			 50.91%±9.543 [47.494, 54.323]	 60.53%±26.043 [51.213, 69.852]	 (56.71%)	 (98.11%)

Table 5.2 (*continued*).

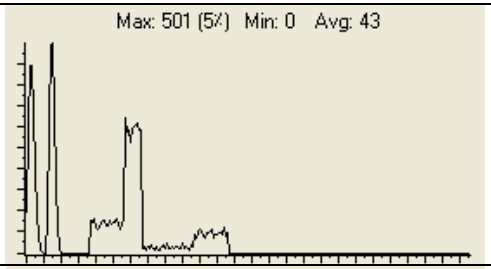
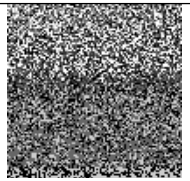
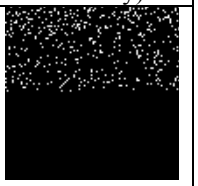
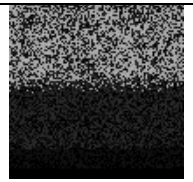
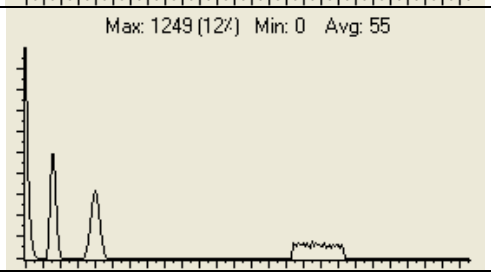
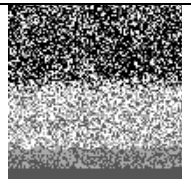
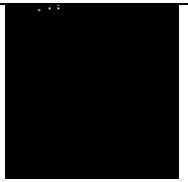
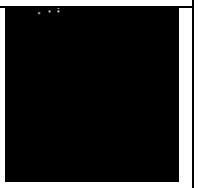
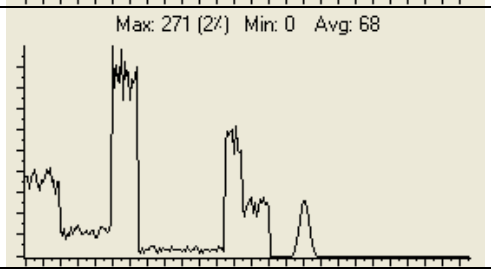
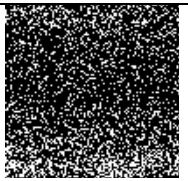
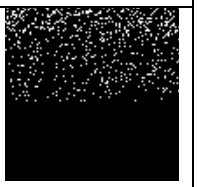
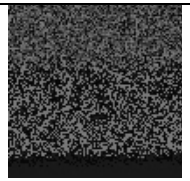
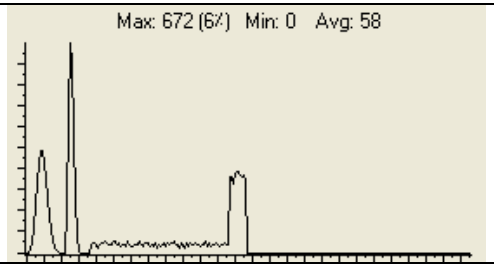
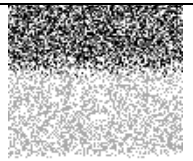
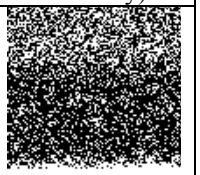
Bench Mark No.	Synthetic Image	Histogram	K-Means Sample TM_c (Avg. Classification Accuracy±SD) [CI]	PSO Sample TM_c (Avg. Classification Accuracy±SD) [CI]	Best K-Means TM Difference (Best Classification Accuracy)	Best PSO TM Difference (Best Classification Accuracy)
7			 60.62%±27.284 [50.856, 70.383]	 77.14%±14.021 [72.119, 82.154]	 (95.81%)	 (95.59%)
8			 99.96%±0 [99.96, 99.96]	 99.96%±0 [99.96, 99.96]	 (99.96%)	 (99.96%)
9			 55.23%±20.975 [47.723, 62.734]	 60.19%±13.428 [55.384, 64.995]	 (75.99%)	 (94.08%)

Table 5.2 (*continued*).

Bench Mark No.	Synthetic Image	Histogram	K-Means Sample TM_c (Avg. Classification Accuracy±SD) [CI]	PSO Sample TM_c (Avg. Classification Accuracy±SD) [CI]	Best K-Means TM Difference (Best Classification Accuracy)	Best PSO TM Difference (Best Classification Accuracy)
10			 88.83%±0 [88.83, 88.83]	 60.14%±0.211 [60.065, 60.216]	 (88.83%)	 (60.47%)
		Average Performance	81.74%±18.25123 [70.43,93.05]	83.73%±16.63667 [73.41,94.04]		