CHAPTER 2

Principles and Theories of the Study

This chapter describes some principles that will be used in this research. The sections start with introducing the fundamental theory of a fuzzy set and fuzzy clustering algorithms. Then, gray level co-occurrence matrix (GLCM) is presented.

2.1 Fuzzy set theory

Zadeh [32 - 33] introduced the theory of fuzzy set [32 – 41] in 1965. The fuzzy set defines an element by the membership values in the interval [0, 1] while the crisp set defines the membership values in {0, 1}. This mechanism can be used to describe various information, both quantitative and qualitative. Therefore, fuzzy set is widely used in many areas such as in medical field, military, and industrial field. Fuzzy set is formally defined as follows:

A fuzzy set is characterized by a membership function mapping the element of a domain, space, or universe of discourse **X** to the unit interval [0, 1] [36]. It can be written in mathematical formula as follows:

$$A: \mathbf{X} \rightarrow [0, 1]. \tag{2.1}$$

For the discrete and finite universe **X**, the fuzzy set can be written in a form of summation as follows [7, 9]:

$$A = \frac{a_1}{x_1} + \frac{a_2}{x_2} + \ldots + \frac{a_n}{x_n} = \sum_{i=1}^{n} \frac{a_i}{x_i}. \tag{2.2}$$

Note that the summation and division are not an algebraic notation. When the universe **X** is continuous, the fuzzy set can be represented with the integral symbolic as

$$A(x) = \int_x \frac{a}{x}. \tag{2.3}$$

Again, this integral is not an algebraic notation. To use the fuzzy set, the problem is how to select the membership function. There is no unique solution. A user should select or create a suitable function for each problem. Equations (2.4) – (2.11) show some of the popular membership functions.

1. Triangular function:

$$A(x) = \begin{cases} 0, & if\ x \le a \\ \dfrac{x-a}{m-a}, & if\ x \in [a, m] \\ \dfrac{b-x}{b-m}, & if\ x \in [m, b] \\ 0, & if\ x \ge b, \end{cases} \tag{2.4}$$

where m is a modal value, a and b denote the lower and upper bounds, respectively, for nonzero values of $A(x)$.

2. Γ- function:

$$A(x) = \begin{cases} 0, & if\ x \le a \\ 1 - e^{-k(x-a)^2}, & if\ x > a \end{cases}, \text{ where } k > 0 \tag{2.5}$$

or

$$A(x) = \begin{cases} 0, & if\ x \le a \\ \dfrac{k(x-a)^2}{1-k(x-a)^2}, & if\ x > a \end{cases}, \text{ where } k > 0. \tag{2.6}$$

3. S – function:

$$A(x) = \begin{cases} 0, & if\ x \le a \\ 2\left(\dfrac{x-a}{b-a}\right)^2, & if\ x \in [a, m] \\ 1 - \left(\dfrac{x-b}{b-a}\right)^2, & if\ x \in [m, b] \\ 1, & x > b \end{cases} \tag{2.7}$$

point $m = (a - b)^2$ is known as a crossover of the S-function.

4. Trapezoidal function:

$$A(x)=\begin{cases}0, & if\ x<a\\ \dfrac{x-a}{m-a}, & if\ x\in[a,m]\\ 1, & if\ x\in[m,n]\\ \dfrac{b-x}{b-n}, & if\ x\in[n,b]\\ 0, & if\ x>b\end{cases}. \tag{2.8}$$

5. Gaussian function:

$$A(x)=e^{-k(x-m)^2},\ \text{where } k>0. \tag{2.9}$$

6. Exponential-like function:

$$A(x)=\frac{1}{1+k(x-m)^2},\ \text{where } k>1 \tag{2.10}$$

or

$$A(x)=\frac{k(x-m)^2}{1+k(x-m)^2},\ \text{where } k>0. \tag{2.11}$$

2.2 Fuzzy clustering

Clustering is an unsupervised method that can be used to classify data into groups based on the similarity among the individual data items. Typically, we do not know the exact number of clusters. Clustering algorithm estimates the number of clusters by calculating from the cluster density and distribution. Moreover, the cluster distribution may include size, shape, and prior knowledge. Cluster will be divided into two clusters when the cluster density is small and the cluster distribution is very high. For the cluster members, the membership grade is normally computed from the normalized distance between the data item to the cluster prototypes where the cluster prototypes are the cluster centers [36, 42].

In this section, we introduce the notation of data set, cluster prototypes, fuzzy partition, and a Fuzzy *C*-Means (FCM) clustering [42].

2.2.1 The data set

Typically, the data set [36] for clustering can be a quantitative (numerical), a qualitative (categorical), or a mixture of both. However, we explain only quantitative data that can be found in a general application.

Let $\mathbf{X} = \{\mathbf{x}_j \mid j = 1, 2, \ldots, N\}$, $\mathbf{x}_j = [x_{1j}, \ldots, x_{nj}]^T$, and $\mathbf{x}_j \in \Re^n$. Then, $\mathbf{X}$ is represented as an n x N matrix as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix}. \tag{2.12}$$

In the pattern recognition terminology, the columns of this matrix are called patterns or objects. The rows are called the features or attributes, and $\mathbf{X}$ is called the pattern or data matrix. The meaning of the columns and rows of $\mathbf{X}$ depends on the context.

2.2.2 Cluster prototypes

To reduce the computation time, the clustering algorithms compute the similarity between the individual data vector and the cluster prototype [42]. The cluster prototypes are not known before applying the clustering technique but it is assigned during initial clustering process. It can be similar to one of the member of the group or the mean of the cluster data. Prototype in each cluster can be a single prototype or multi-prototype depending on the characteristic of the data set or the application requirement.

2.2.3 Fuzzy partition

Related to the fuzzy set, the fuzzy partition [42] allows μ_{ij} into a real interval values [0, 1] where μ_{ij} is the membership grade of feature vector x_j in cluster i. A fuzzy partition is represented by the fuzzy partition matrix $\mathbf{U} = \left[\mu_{ij}\right]_{c \times N}$. The fuzzy partition matrix conditions are as follows:

$$\mu_{ij} \in [0,1],\ 1 \le i \le c,\ 1 \le j \le N, \tag{2.13}$$

$$\sum_{i=1}^{c} \mu_{ij} = 1,\ 1 \le j \le N, \tag{2.14}$$

$$0 \le \sum_{j=1}^{N} \mu_{ij} < N,\ 1 \le i \le c\,. \tag{2.15}$$

The i^{th} row of the fuzzy partition matrix **U** contains the value of the i^{th} membership function of the fuzzy subset $\mathbf{A}_i$ of **X**. Equation (2.14) constrains the sum of each column to 1, and thus the total membership of each $\mathbf{x}_j$ in **X** equals one. The fuzzy partitioning space for **X** is represented as

$$M_{fc} = \left\{ U \in \Re^{c \times N} \middle| \mu_{ij} \in [0,1], \forall i, j; \sum_{i=1}^{c} \mu_{ij} = 1, \forall j; 0 \le \sum_{j=1}^{N} \mu_{ij} < N, \forall i \right\}. \tag{2.16}$$

2.2.4 The Fuzzy *C*-Means (FCM) clustering

The FCM clustering [42] assumes that the clusters are approximately the same size in spherical space. The objective function for FCM is formulated as

$$J(X;U,V) = \sum_{i=1}^{c} \sum_{j=1}^{N} \left(\mu_{ij}\right)^{m} \left\| \mathbf{x}_j - \mathbf{v}_i \right\|_{\mathbf{A}}^{2}, \tag{2.17}$$

where the parameters **U**, **V**, $D_{ij\mathbf{A}}^{2}$, and *m* are a fuzzy partition matrix of **X**, a set of vector of cluster prototypes, a squared inner-product distance norm, and a fuzzifier, respectively. The value of the cost function (2.17) can be seen as a measure of the total variance of $\mathbf{x}_j$ from $\mathbf{v}_i$. These parameters are formulated as follows:

$$\mathbf{U} = \left[\mu_{ij}\right] \in M_{fc}, \tag{2.18}$$

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c], \mathbf{v}_i \in \Re^{p}, \tag{2.19}$$

$$D_{ij\mathbf{A}}^{2} = \left\| \mathbf{x}_j - \mathbf{v}_i \right\|_{\mathbf{A}}^{2} = \left(\mathbf{x}_j - \mathbf{v}_i\right)^{T} \mathbf{A} \left(\mathbf{x}_j - \mathbf{v}_i\right), \tag{2.20}$$

$$m \in [1, \infty). \tag{2.21}$$

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be a set of vectors, where each vector is an *N*-dimensional vector. Choose the number of clusters $1 < c < n$. The update equations for FCM [42] are as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{D_{ij\mathbf{A}}}{D_{kj\mathbf{A}}}\right]^{\frac{2}{m-1}}}, 1 \le i \le c, 1 \le j \le N, \tag{2.22}$$

$$\mathbf{v}_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{m} \mathbf{x}_j}{\sum_{j=1}^{N} \mu_{ij}^{m}}, 1 \le i \le c. \tag{2.23}$$

The following process is the summarization of the FCM clustering algorithm [42].

```
Fix the number of clusters c
Initiate prototypes
Do {
    Update membership using (2.22)
    Update prototypes using (2.23)
} Until prototypes stabilize
```

2.3 Gray Level Co-occurrence Matrix (GLCM)

A gray level co-occurrence matrix [4, 7] was proposed by Haralick in 1972 [4]. It is a second-order statistics of an image. GLCM represents the cardinality of two gray levels i and j at a desire distance (d) and orientation (θ). The orientation assignment of GLCM is shown in figure 2.1. The joint probability of occurrence of two gray level values is counted as a $P(i, j; d, \theta)$. Suppose the size of an image is $N_x \times N_y$. Let $L_x = \{1, 2, \ldots, N_x\}$ and $L_y = \{1, 2, \ldots, N_y\}$ be the horizontal and vertical spatial domain, respectively, and $G = \{1, 2, \ldots, N_g\}$ be the set of N_g quantized gray tones. The image I can be represented as a function which assigns some gray tones in G to each resolution cell or a pair of coordinates in $L_y \times L_x$.

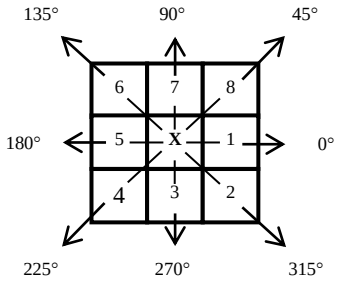


Figure 2.1 The GLCM orientation assignment.

The joint probability of occurrence of two gray level values in each distance and direction is calculated as:

$$P(i, j, d, 0) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = 0, |l - n| = d, I(k, l) = i, I(m, n) = j\}$$

$$P(i, j, d, 45) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = d, |l - n| = -d, I(k, l) = i, I(m, n) = j\}$$

$$P(i, j, d, 90) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = d, |l - n| = 0, I(k, l) = i, I(m, n) = j\}$$

$$P(i, j, d, 135) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = d, |l - n| = d, I(k, l) = i, I(m, n) = j\} \quad (2.24)$$

$$P(i, j, d, 180) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = 0, |l - n| = -d, I(k, l) = i, I(m, n) = j\}$$

$$P(i, j, d, 225) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = d, |l - n| = -d, I(k, l) = i, I(m, n) = j\}$$

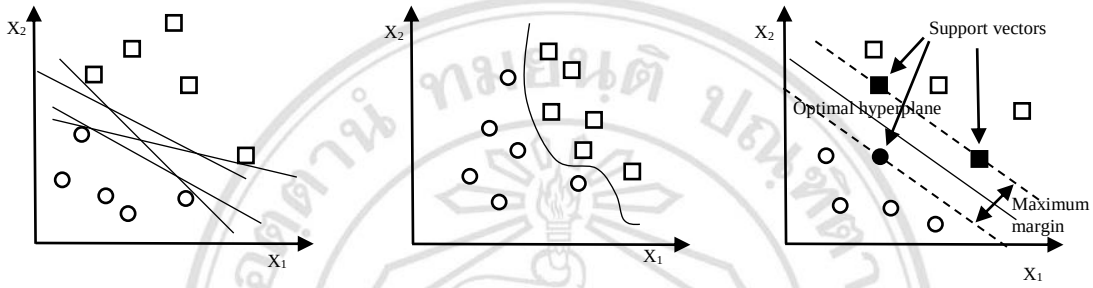
$$P(i, j, d, 270) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = d, |l - n| = 0, I(k, l) = i, I(m, n) = j\}$$

$$P(i, j, d, 315) = \#\{((k, l), (m, n)) \in (L_y \times L_x) \times (L_y \times L_x) \mid |k - m| = d, |l - n| = d, I(k, l) = i, I(m, n) = j\}$$

where # denotes the number of elements in the set.

2.4 Multi-class Support Vector Machine (MSVM)

Support Vector Machine (SVM) [43-44] can be applied as a classifier based on discriminative hyperplane. An example of discriminative hyperplanes is shown in figure 2.2. The plane separates between different class objects. The SVM algorithm finds this plane that has the maximum margin of the training samples as shown in figure 2.2 (c).



(a) Discriminative line (b) Discriminative curve (c) Optimal hyperplane

Figure 2.2 An example of discriminative hyperplanes.

Most classification problems are normally nonlinear problem. Therefore, SVMs are developed with kernel functions to solve the problem. Some of kernels that can be used are shown in table 2.1.

Table 2.1 An example of SVM kernels.

Kernel name	Kernel function
Linear	$\mathbf{X}_i \bullet \mathbf{X}_j$
Polynomial	$\left(\gamma \mathbf{X}_i \bullet \mathbf{X}_j + C\right)^d$
Radial Basis Function (RBF)	$\exp\left(-\gamma \left\lvert \mathbf{X}_i \bullet \mathbf{X}_j \right\rvert^2\right)$
Sigmoid	$\tanh\left(\gamma \mathbf{X}_i \bullet \mathbf{X}_j + C\right)$

where $K\left(\mathbf{X}_i, \mathbf{X}_j\right) = \phi\left(\mathbf{X}_i\right) \bullet \phi\left(\mathbf{X}_j\right)$. Input data points $\mathbf{X}_i$ are transformed by mapping into the higher dimensional feature space $\phi\left(\mathbf{X}_i\right)$. Gamma is an adjustable parameter of

certain kernel functions. Because of their localized and finite responses across the entire range of the real x-axis, the RBF is the most popular choice of kernel types used in SVM.

The SVM used in the experiment is the one with soft margin optimization. The Radial Basis Function (RBF) used in each SVM is

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{2\sigma^2} \right). \tag{2.25}$$

For the unbias comparisons of experimental results, we used the multi-class support vector machine (MSVM) [43] as classifier. In this research, we used one-versus-all strategy. The MSVM is a method that assigns a class label to a vector as one of the several classes. Suppose we have an optimum discriminant function ($D_i(\mathbf{x})$ for $i = 1, \ldots, C$). From the support vector machine (SVM) [44], we have an optimum hyperplane at $D_i(\mathbf{x}) = 0$ that will separate class i from all the others. Hence, each SVM classifier gives $D_i(\mathbf{x}) > 0$ for vectors in class i, and $D_i(\mathbf{x}) < 0$ for those in all other classes. Then, the classification rule is

$$\mathbf{x} \text{ is assigned to class } i \text{ if } i = \underset{j=1,\ldots,n}{\operatorname{argmax}} D_j(x). \tag{2.26}$$