2. MODIFIED ALGORITHMS FOR CLUSTERING MULTISPECTRAL REMOTELY SENSED DATA

"Every duty which is bidden to wait, returns with seven fresh duties at its back."
- Charles Kingsley

2.1 Introduction

Clustering is the process of obtaining natural classes without any priori knowledge of prototype classification. The purpose of clustering is to group together (aggregation) data points which are close to one another. In other words, clustering is viewed as a process of partitioning (division) data set into subsets or clusters of samples such that the degree of similarity is high among members of the same cluster and low between members belonging to different clusters.

A variety of cluster seeking algorithms are available in literature. Broadly they can be classified as: hierarchical methods (Agglomerative and Divisive) and non-hierarchical methods (direct and indirect). A hierarchical clustering technique imposes a hierarchical structure on the data which consists of a sequence of clustering. The sequence of forming groups proceeds such that whenever two

* Some parts of the material in this chapter appeared in the following research papers:


samples belong (do not belong) to the same cluster at some level, they remain together (separated) at all the subsequent levels. The hierarchy is decided by the linkage method viz., single linkage (nearest neighbour), complete linkage (farthest neighbour), group average link, weighted average link, centroid and median method.

Depending on the algorithmic approach taken, a hierarchical structure begins with $n$ clusters, one per pattern, and grows a sequence of clustering until all $n$ patterns are in a single cluster (Agglomerative approach). Successive merging operation is the heart of Agglomerative methods. Merging is the process of gathering together, on the basis of a similarity measure, two samples and assigning them same cluster membership or label so that the pair becomes a building block for further clustering.

Divisive methods employ successive splitting operation. Division procedures start with the entire set of samples in one class and the set of $n$ patterns are divided into smaller clusters according to some similarity or dissimilarity measure. These subsets are further subdivided and the process is continued till a suitable stopping rule arrests further subdivisions.

Unlike hierarchical methods, which give a sequence of partitions, the essence of a non-hierarchical method is to choose some initial partition of the data set and then alter cluster memberships so as to obtain a better partition according to some objective function. Non-hierarchical algorithms are characterized as direct method if no criterion functions are used to optimize the performance of the classification. On the other hand, indirect methods are based upon certain criterion functions (also known as performance indices), which are optimized (maximized or minimized) to get the best fit of data into categories. Some of these non-hierarchical algorithms are discussed in Chapter 4.
2.2 Review of Existing Methodologies

There exist a variety of hierarchical clustering techniques in literature (Duda and Hart, 1973; Jain and Dubes, 1988; Anil Jain, 1995). Traditional hierarchical Agglomerative and Divisive clustering algorithms (Gowda and Krishna, 1977, 1978b; Takio Kurita, 1991; Shi Xu et al., 1993; Zhang et al., 1994) can be applicable to classify only small data sets.

Takio Kurita (1991) introduces the heap strategy in carrying out the Agglomerative clustering algorithm and provides a comparative assessment of computational time requirements of his method over the straight forward Agglomerative method.

The conventional nearest neighbourhood considers only nearness as the criterion for neighbourhood. Given a set of points in multi-dimensional space, Chaudhuri (1995) proposed an iterative procedure to compute the neighbours of an arbitrary point \( p \), so that the neighbours should be as near to \( p \) and as symmetrically placed around \( p \) as possible. Amadasun and King (1988) used uniform neighbourhoods and uniform criterion for segmentation of multispectral images via Agglomerative clustering.

Perhaps a more natural way to define clusters is by utilizing the property of mutual nearest neighbourhood between two patterns (Gowda, 1978a; Nagabhushan, 1988; Zhang, 1994). Gowda and Krishna (1977, 1978a, 1978b) suggested a method for determining the Mutual Nearest Neighbours (MNN) concept for Agglomerative and Disaggregative clustering. In their work, a two-way nearness value MNV (Mutual Neighbourhood Value) of a sample point, using the conventional nearest neighbours, is considered to agglomerate the samples. A pragmatic stopping rule is proposed in (Shi Xu et al., 1993) to select good partitions using Agglomerative hierarchical clustering procedures. This rule combines the distance dissimilarity in a dendrogram with the squared-error criterion for the evaluation of partitions in all
hierarchical levels.

Spanning trees are a combinational way to introduce hierarchical clustering and to provide efficient algorithms to build clusters optimizing some criterion. In his paper, Nagabhushan (1988) defined mutual neighbour distance \((mnd)\) and discussed the superiority of this measure over MNV. He uses this index in a graph theoretic environment to build a minimal spanning tree.

The mutual nearest neighbourhood concept introduced by Gowda (1978a, 1978b) was used by Zhang et al., (1994) for data clustering by using the concept of psychological potential field. These efficient algorithms are attempted to reduce the computational time and memory requirements. However, no completely satisfactory method is found to suit for multispectral data classification. Motivated by this, the concept of mutual nearest neighbourhood is further extended in this work to classify huge volume remotely sensed data more efficiently.

Due to multidimensionality of satellite data, the classification time increases drastically depending on image size and vice versa. The limitation encountered in applying the conventional Agglomerative and Divisive clustering is, computational time is very high in finding \(k\) nearest neighbours that are necessary for initiating merging or splitting operations. In order to minimize the time and space requirements, a new concept is proposed, namely, Area of Interest around each pixel.

Distance measures are important issues in applied science and engineering problems, particularly in Pattern Recognition and Data Analysis. The thrust of research in the area of cluster analysis has been towards determining various object similarity or proximity measures, and developing clustering algorithms utilizing them. A large number of such measures and corresponding clustering methods have been developed to date. Comprehensive surveys can be found in (Duda and Hart, 1973; Jain and Dubes, 1988). In all these methods, the task is to determine clusters, such that objects in the same cluster exhibit a high degree of similarity.
(high intra-cluster similarity), while objects from different clusters exhibit a low
degree of similarity (low inter-cluster similarity). So, the effectiveness of any
clustering method depends, to a great extent on the distance measures adopted.

Minkowski metrics (Duda and Hart, 1973; Jain and Dubes, 1988) are often used
as convenient dissimilarity measures in classification methods and cluster analysis
if the objects are described with respect to quantitative feature variables. When
each object is described by qualitative feature on the other hand, for example
binary variables, various distance measures (eg., Hamming, Russel and Rao,
Jaccard and Needham, etc.,) are employed (Duda and Hart, 1973). Since the
conventional data space is Euclidean, it is natural to use Euclidean distance in such
a space. To reduce the computations, City-block, Chess board and other distances
are often used in Image Processing and related problems.

In determining similarity and dissimilarity between pixels, only the feature
values of pixels are employed by these traditional distance measures. Even though
these distances are computationally efficient, one encounters many pixels having
the same proximity value with many other pixels when the $k$-width neighbourhood
is considered in remotely sensed data. Hence, this leads to conflict in choosing $k$
nearest neighbour pixels, causing the accuracy of the final result to suffer (cause
for this is given in the sequel). Clustering procedures that employ such distance
measures give erroneous classification. To alleviate this problem, a modified
distance measure is proposed which takes into account the relative position of the
pixels in coordinate space along with the feature spectral values or band values of
the pixels.

The present work suggests a new avenue by which Agglomerative and
Disaggregative clustering using the concept of mutual nearest neighbourhood can
be applied to remotely sensed data for quantitative analysis, so that the intrinsic
limitations of applying these procedures to data of huge volume can be overcome.
2.3 New Concepts Employed

2.3.1 Notion of Area of Interest

In traditional Agglomerative and Divisive clustering procedures, the computational time and memory requirements are high as the whole image is considered to determine the $k$ nearest neighbourhood for each pixel necessary for initiating merging or splitting operation respectively. To realize reduction of memory and computational time, an Area of Interest is defined, around each pixel to search the required $k$ nearest neighbours.

While working on an image data for classification, some basic knowledge like, roughness or smoothness of image areas is necessary. Image areas of high frequency are tonally 'rough', i.e., the grey levels in these areas change abruptly over a relatively small number of pixels (eg., across roads or field borders). 'Smooth' image areas are those of low frequency (eg., forest areas or water source), their grey levels vary gradually over a relatively large number of pixels. Such a situation is very common in remotely sensed data, where the next pattern to be classified is frequently close to the one just encountered. Thus, the density of the neighbouring pixels belonging to the same class is high, i.e., pixels are auto-correlated.

In defining the Area of Interest with reference to the pixel under consideration, an assumption is made that the probability of existence of the $k$ nearest neighbours surrounding the given pixel is very high (as mentioned earlier). Hence, the $k$ nearest neighbours will be effectively selected by considering the Area of Interest around each pixel.

Area of Interest is specified by the user as an area in the coordinate space surrounding each pixel under consideration. All pixels falling within this Area of Interest are then used for calculating the distance from the pixel under
consideration for finding $k$ nearest neighbours. Consequently, the $k$-width neighbourhood pixels within this Area of Interest are then considered for further merging or splitting operations.

In Fig. 2.1, the dotted line shows the Area of Interest for a given pixel $p$ in the image frame shown by the thick boundary. Let $h$ and $w$ be the users specified height and width of the Area of Interest, and $a$, $b$, $c$, and $d$ represent the corner points. Let $(x, y)$ be the coordinate vector of pixel $p$. The coordinate vectors of corner points of the Area of Interest are calculated by the following equations.

$$
\begin{align*}
    a(x', y') &= ((x - \frac{w}{2}), (y + \frac{h}{2})) \\
    b(x', y') &= ((x + \frac{w}{2}), (y + \frac{h}{2})) \\
    c(x', y') &= ((x - \frac{w}{2}), (y - \frac{h}{2})) \\
    d(x', y') &= ((x + \frac{w}{2}), (y - \frac{h}{2}))
\end{align*}
$$

Figure 2.1 Area of Interest in image space
Only those pixels present in this window are considered for finding $k$ nearest neighbours for a given pixel $p$. The Area of Interest is used also to emphasize or de-emphasize degree of nearness of neighbourhood pixels present in the vicinity of this area and it deals with the population present within this area rather than with the whole image.

2.3.2 Proposed Modified Distance Measure

In Agglomerative and Divisive clustering method, when $k$ nearest neighbourhood is considered in multispectral image data, one encounters many pixels having the same proximity value with many other pixels. This leads to conflict in choosing $k$ nearest neighbour pixels causing erroneous classification, since many pixels may not be involved in the merging or splitting operations. For example, consider Fig. 2.2, which shows nearest neighbours for a pixel '+' in the image area.

Let there be $m$ number of pixels '*' having the same proximity values with respect to the pixel '+' This leads to conflict in choosing required $k$ nearest neighbours when $k < m$. Thus, a new concept is introduced for finding $k$ nearest neighbours, by selecting $k$ nearest neighbours surrounding the pixel '+' enclosed with in the Area of Interest, shown by dotted window in Fig. 2.2. The proposed modified distance measure incorporates this concept.

![Figure 2.2 Nearest neighbours for pixel '+']
The modified distance measure fuses the relative positions of the pixels with their feature values. However, it may be noted that traditional distance measures consider only feature values of the pixels but not their relative positions. Figure 2.3a represents the feature space for pixels $a$ and $b$ having the feature vectors with $D$ number of features $F_a (f_1, f_2, \ldots, f_D)$ and $F_b (f_1', f_2', \ldots, f_D')$ respectively. Figure 2.3b represents the coordinate space for pixels $a$ and $b$ having coordinate vectors $C_a (x, y)$ and $C_b (x', y')$ respectively. The distance $d(a, b)$ between pixels $a$ and $b$ is given by

$$d(a, b) = d_1(F_a, F_b) + d_2(C_a, C_b)$$

(2.5)

where $d_1(F_a, F_b)$ is a function of any distance measure $^\dagger$ between pixels $a$ and $b$ based on their feature vectors, $d_2(C_a, C_b)$ is the distance between pixels $a$ and $b$ with respect to their relative positions (coordinate vectors).

$^\dagger$ as found in any traditional distance measures like euclidean distance, cityblock distance etc.
Let $d_1(F_a, F_b)$ be the Euclidean distance between pixels $a$ and $b$ based on their feature vectors $F_a$, $F_b$ given by

$$
d_1(F_a, F_b) = A(\sqrt{\sum_{i=1}^{D} (F_{a_i} - F_{b_i})^2})
$$

(2.6)

Here $A$ is a function that converts real proximity value to integer value, such as $A(6.2345678) = 62345$ (proximity value is converted into integer number up to $4^{th}$ decimal place).

In equation (2.5), $d_2(C_a, C_b)$ is the distance between pixels $a$ and $b$ based on their coordinate vectors $C_a$, $C_b$ and is given by

$$
d_2(C_a, C_b) = G(|x - x'|, |y - y'|) / G(X, Y)
$$

(2.7)

Here $X$ and $Y$ are width and height of the image respectively, and $G$ is a function which finds the greatest of two numbers. For example, $G(4, 6) = 6$.

Consider two pixels $A$ and $B$ having the feature vectors with three features $(20, 22, 105)$ and $(46, 32, 165)$ respectively and let their coordinate vectors be $(10, 20)$ and $(30, 22)$ respectively. Let the image size be $100 \times 100$. Then, from the equation (2.5) $d(A, B) = 661513 + 0.20 = 661513.20$.

It may be noted that the modified distance measure is of structure $d_1.d_2$, where $d_1$ ($d_1 \geq 0$) is the integral and $d_2$ ($d_2 < 1$) is the fractional part of the modified distance $d$. Even though, $d_1 >> d_2$, distance $d_2$ helps to achieve better discrimination of pixels as it is calculated based on the relative positions of the pixels. Thus, when modified distance measure is employed, in most of the cases no two identical pixels are associated with the same proximity value.
2.4 Computational Algorithms

The proposed Agglomerative and Disaggregative clustering algorithms are nonparametric and hierarchical and are based on modified distance measure. The criterion is to merge the pixels based on MNV (Mutual Neighbourhood Value) (Gowda, 1977; Nagabhushan, 1988). In this algorithm, a mutual neighbourhood value (MNV) between two pixels is measured. When $p_A$ is the $s^{th}$ nearest neighbour of $p_B$, and $p_B$ is the $t^{th}$ nearest neighbour of $p_A$, then the MNV between $p_A$ and $p_B$ is $\text{MNV}(A, B) = \text{MNV}(B, A) = s + t$. The smaller the MNV, the more similar the pixels are. The minimum MNV is $1 + 1 = 2$ and the maximum MNV is $2k$ ($k$ is the neighbourhood depth).

For illustration, consider pixels $p_A, p_B, p_C, p_D, p_E$ in an image frame as shown in Fig. 2.4. The numbers in the figure are the distances between pairs of pixels.

![Figure 2.4 An example of five pixels in the image space](image)

The 3-nearest neighbour for every pixel is decided and is given in Table 2.1. In row 1, $p_B$ is the first nearest neighbour, and $p_C$ and $p_E$ are the second and third nearest neighbours of $p_A$. For $p_E$ in row 5, the first, second and third nearest neighbours are $p_B, p_D$ and $p_C$ respectively. The number within the parenthesis under each neighbour pixel indicates the MNV between the pixel and the first pixel in that row. For example, MNV between $p_A$ and $p_B$ is $(1 + 1) = 2$. According
Table 2.1 MNVs for the five pixels

<table>
<thead>
<tr>
<th></th>
<th>1st NN</th>
<th>2nd NN</th>
<th>3rd NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_A$</td>
<td>$p_B$</td>
<td>$p_C$</td>
<td>$p_E$</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(100)</td>
</tr>
<tr>
<td>$p_B$</td>
<td>$p_A$</td>
<td>$p_E$</td>
<td>$p_C$</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(5)</td>
</tr>
<tr>
<td>$p_C$</td>
<td>$p_A$</td>
<td>$p_B$</td>
<td>$p_D$</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(5)</td>
<td>(4)</td>
</tr>
<tr>
<td>$p_D$</td>
<td>$p_C$</td>
<td>$p_E$</td>
<td>$p_B$</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(4)</td>
<td>(100)</td>
</tr>
<tr>
<td>$p_E$</td>
<td>$p_B$</td>
<td>$p_D$</td>
<td>$p_C$</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(4)</td>
<td>(100)</td>
</tr>
</tbody>
</table>

to row 5, $p_A$ is not among 3-nearest neighbours of $p_E$. Thus, $p_A$ and $p_E$ are not mutual neighbours considering a neighbourhood depth of $k = 3$. A large number, 100 ($>2k$), is written below $p_E$ in row 1.

2.4.1 Agglomerative Clustering Procedure

The mutual nearest neighbourhood based Agglomerative clustering algorithm proceeds as follows:

1. Let $P = \{\{p_{11}, p_{12}, p_{13}, \ldots, p_{1X}\}, \{p_{21}, p_{22}, p_{23}, \ldots, p_{2X}\}, \ldots, \{p_{Y1}, p_{Y2}, p_{Y3}, \ldots, p_{YX}\}\}$ be a set of $X$ by $Y$ pixels on $D$ dimensions. Let the initial number of clusters be $X$ by $Y$ with each cluster having a cluster weight of 1.

2. Find the $k$-width neighbourhood pixels for each pixel by following steps from 2(a) to 2(c).

2(a). Compute the coordinates of the boundary corners of Area of Interest using equation (2.1) to (2.4).
2(b). Compute proximity values using modified distance measure given by the eq. (2.5), between all pairs of pixels coming under the Area of Interest.

2(c). Search the $k$-width nearest neighbours having the least proximity values.

3. Compute the MNV for every pair of pixels. If $P_i$ (1 ≤ $i$ ≤ $X$ by $Y$) and $P_j$ (1 ≤ $j$ ≤ $k$) are not mutual neighbours for the $k$, set MNV($i, j$) to arbitrarily large number such as 100 ( > $2k$).

4. Set MNV threshold MNV$_a$ = 2.

5. Identify all the pairs of pixels with MNV = MNV$_a$. Merge every such pair into a cluster in ascending proximity values, and reduce the total number of clusters by 1 after each merges.

6. If all the pairs of pixels with MNV = MNV$_a$ have been merged and if still MNV$_a$ < $2k$, then MNV$_a$ = MNV$_a$ + 1, go to step 5. Otherwise, stop.

2.4.2 Disaggregative Clustering Procedure

The mutual nearest neighbourhood based Disaggregative clustering algorithm proceeds as follows:

1. Let $P = \{(p_{11}, p_{12}, p_{13}, ..., p_{1X}), (p_{21}, p_{22}, p_{23}, ..., p_{2X}), ..., (p_{Y1}, p_{Y2}, p_{Y3}, ..., p_{YY})\}$ be a set of $X$ by $Y$ pixels on $D$ dimensions. Let the initial number of clusters be 1 having a cluster weight of $X$ by $Y$.

2. Find the $k$-width neighbourhood pixels for each pixel by following steps from 2(a) to 2(c).

2(a). Compute the coordinates of the boundary corners of Area of Interest using equation (2.1) to (2.4).

2(b). Compute proximity values using modified distance measure (given in eq. 2.5), between all pairs of pixels coming under the Area of Interest.
2(c). Search the $k$-width nearest neighbours having the least proximity values.

3. Compute the MNV for every pair of pixels. If $P_i (1 \leq i \leq X$ by $Y)$ and $P_j$(1 $\leq j \leq k)$ are not mutual neighbours for the $k$, set MNV$(i, j)$ to arbitrarily large number such as 100 ($> 2k$).

4. Let $T_{merge}$ and $T_{split}$ ($\leq 2k$) be the threshold of the MNV chosen for carrying out the Disaggregative process. Strong and weak clusters will be obtained depending on the value of $T_{merge}$ and $T_{split}$.

5. Identify the pairs of pixels with MNV = $T_{split}$ in set $P$ having descending proximity values. Split such pairs into a separate clusters $C_i$ and $C_j$, and increase the total number of clusters by 2.

6. Find all the mutual nearest neighbours (with MNV $\leq T_{merge}$) of pixel in $C_i$ and pixel in $C_j$. Add them to the corresponding $C_i$ and $C_j$ clusters and find all its mutual nearest neighbours (with MNV $\leq T_{merge}$). Out of them assign all those that are not already present in $C_i$ to $C_i$ and $C_j$ to $C_j$ and delete them from the set $P$. Repeat this procedure for the third and subsequent elements in $C_i$ and $C_j$.

7. If all the pairs of pixels with MNV = $T_{split}$ are splitted, then make $T_{split}$ = $T_{split}$ - 1 and repeat steps 5 and 6 until the set $P$ is not empty. Else repeat steps 5 and 6 until the set $P$ is not empty.

### 2.5 Experimentation

Experiments have been conducted to demonstrate the capability and applicability of the proposed clustering methodologies on multispectral remotely sensed data sets. Following experiments are carried out and the results are given below. The results of conventional K-means algorithm are also presented to make a comparative study. The display scheme illustrated in Appendix A.3 is employed in this chapter to plot the output classification maps.
Experiment No. 1:

The first experiment is conducted on the Gaussian generated data set. The objects of numeric type were drawn from a mixture of normal distributions with known number of classes and classification labels so that the results validate the efficacy of the algorithm for clustering the objects and finding the number of samples in each class. The test set is drawn from a mixture of $C$ normal distributions with mean $m_i$ and covariance matrix $COV_i$ having individual variances of 3.87 and zero covariance. The numbers of samples generated in each class and their mean values chosen are shown in Table 2.2. The test samples were independently generated using Gaussian vector generator.

The application of the proposed Agglomerative algorithm, for a neighbourhood width of $k = 18$ and $w = h = 10$, and the Disaggregative algorithm, for a neighbourhood width of $k = 20$ and $w = h = 10$ yielded six classes. The classification result obtained from the conventional K-means algorithm is also presented to compare the classification results of the proposed algorithms.

As indicated in Table 2.2, there is a perfect agreement between the number of classes generated and the number of clusters predicted by the algorithms. In all the six classes, the classification results were almost in full agreement with the test samples generated.

Experiment No. 2:

The multispectral IRS (Indian Remote Sensing) satellite data covering Kundapur area of Karnataka State, India is used in this experiment for the classification purpose. Figure A.4.3 shows False Colour Composite (FCC) map of MSS image covering Kundapur area (given in Appendix A.4). This image is of size 31 X 31 with four features. The site consists of a river joining the Arabian Sea and land area mainly made up of different vegetation.
The application of the proposed Agglomerative algorithm, for a neighbourhood width of \( k = 8 \) and \( w = h = 10 \) and the Disaggregative algorithm, for a neighbourhood width of \( k = 12 \) and \( w = h = 10 \) yielded 3 major classes. The classification results from the proposed Agglomerative and Disaggregative methods are given in Fig. 2.5a and Fig 2.5b respectively. The classification result from the K-means algorithm is also presented in Fig. 2.5c as an aid to compare the proposed results. The class legends, their cover types and other related information are tabulated in Table 2.3.

### 2.6 Critical and Comparative Analysis

It is instructive to compare the proposed methods with some other methods to justify the superiority of the proposed methodologies. As a preliminary step, the proposed clustering algorithms are judiciously compared with the existing algorithms found in Gowda and Krishna (1977, 1978b), Takio Kurita (1991), and Zhang et al. (1994) in terms of memory and time requirements. In the following paragraphs, some comparative studies made with other schemes are presented.

Takio Kurita (1991), described an efficient Agglomerative clustering algorithm using a heap strategy. In this paper, Kurita mentioned that the computational time of the Agglomerative algorithm is \( O(N^3) \) when the algorithm is implemented straightforwardly and said that it is applicable only when the numbers of objects are small. He introduced heap concept to speed up the searching process. The computation time to construct a heap of \( N(N-1)/2 \) elements, the case of clustering of \( N \) objects, is \( O(N^2 \log(N)) \). After searching the nearest pairs cluster, the algorithm requires computational time of \( O(\log(N)) \) to update or remove an element from the heap.

Gowda and Krishna (1977, 1978b) proposed hierarchical clustering algorithms based on MNV concept. The clustering algorithms takes computational time of
Modified Algorithms for Clustering Multispectral Remotely Sensed Data

O(N²) and an additional O(Nk) (k-width MNV calculations) for clustering N objects. Whereas, the proposed clustering algorithms given in this chapter require computational time of O(N(w*h)) (w-width and h-height of Area of Interest) and an additional O(Nk) (k-width MNV calculations) for clustering N objects. This shows the computational time required by the proposed clustering algorithms is very less (w*h << N) as compared to the other algorithms stated above.

The clustering algorithm described by Zhang et al. (1994) used MNV concept for clustering data samples. The MNVs of cluster are recalculated during clustering (each merges) which requires more computational time. The clustering algorithm is based on similarity measure. The Eigen vector and Eigen values are used to calculate similarity between two data points and hence naturally demands more computational time.

The algorithm found in Takio Kurita (1991) needs major memory requirements of about 3N² and algorithm found in Gowda and Krishna (1977 and 1978b) requires about 2Nk memory space to store k-width nearest neighbours and k-width MNV values, whereas, the proposed clustering algorithms require almost the same amount of memory (ie., 2Nk) as found in Gowda and Krishna (1977 and 1978b) for clustering N objects.

In their present form, the conventional clustering algorithms are well suited for classifying small data sets and are ascertained from the research papers mentioned in the above paragraphs. When these clustering procedures are applied on multispectral remotely sensed data, one encounters many pixels having the same proximity value with many other pixels as they employ traditional Euclidean and Cityblock distance measures. This leads to conflicts in choosing the k-nearest neighbour pixels causing erroneous classification. It should be noted that it is futile to leave the pixels unlabeled in the classification procedure, since many pixels may not be involved in the merging operation (as MNV > 2k) when the k-nearest neighbourhood is
considered. This problem is alleviated by the modified distance measure proposed in this paper which takes into account the relative position of the pixels in the coordinate space along with the feature values or band values of the pixels. Besides this, strong and weak patterns can be discerned by changing the neighbourhood width $k$, and the algorithms determine the number of clusters depending on this $k$-width.

Further, to authenticate the efficacy of the proposed clustering algorithms, the clusters obtained are examined for their validity using the modified Goodman-Kruskal Gamma ($\gamma$) statistic (Sect. 8.2). The level of significance values obtained using Goodman-Kruskal Gamma ($\gamma$) statistic will indicate the confidence level with which the clusters obtained can be accepted. The validation results are highlighted in Chapter 8 (Sect. 8.5).

### 2.7 Summary

A nonparametric, hierarchical, Agglomerative and Disaggregative clustering algorithms based on mutual nearest neighbourhood concept are proposed for classification of large volume of multispectral data indenting minimum memory and computational time.

To realize the reduction of memory and computational time requirements, an Area of Interest is defined around each pixel to search the required $k$ nearest neighbours necessary for initiating merging or splitting operations. A modified distance measure is proposed, which incorporates the relative position of the pixels with their feature values to improve the performance of the algorithm and reduce the classification error rate. The algorithm is simple, noniterative and can be successfully applied on any image data.

Experiments are conducted to demonstrate the applicability of the proposed methodologies on multispectral images. Further, to authenticate the efficacy of the
proposed algorithms in terms of time and space, these methods are compared with a couple of existing methods. Finally, the clustering results are validated by employing a modified Goodman-Kruskal Gamma (γ) statistic.
2.5a Agglomerative method

2.5b Disaggregative method

2.5c K-means method

2.5a-c Classification map of MSS image near Kundapur, Karnataka, India
Table 2.2 Gaussian generated image details and classification results

<table>
<thead>
<tr>
<th>Class No.s</th>
<th>No. of samples generated</th>
<th>Mean feature values of generating samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>600</td>
<td>220, 150, 15</td>
</tr>
<tr>
<td>2</td>
<td>600</td>
<td>10, 50, 20</td>
</tr>
<tr>
<td>3</td>
<td>600</td>
<td>20, 215, 70</td>
</tr>
<tr>
<td>4</td>
<td>600</td>
<td>50, 220, 50</td>
</tr>
<tr>
<td>5</td>
<td>600</td>
<td>150, 115, 50</td>
</tr>
<tr>
<td>6</td>
<td>600</td>
<td>250, 40, 15</td>
</tr>
</tbody>
</table>

Table 2.2 Continued

<table>
<thead>
<tr>
<th>Class No.s</th>
<th>Proposed Agglomerative method</th>
<th>Proposed Disaggregative method</th>
<th>Conventional K-means method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>598</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>2</td>
<td>598</td>
<td>598</td>
<td>600</td>
</tr>
<tr>
<td>3</td>
<td>597</td>
<td>599</td>
<td>600</td>
</tr>
<tr>
<td>4</td>
<td>600</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>5</td>
<td>600</td>
<td>599</td>
<td>600</td>
</tr>
<tr>
<td>6</td>
<td>598</td>
<td>600</td>
<td>600</td>
</tr>
</tbody>
</table>

Table 2.3 Classification results of major land covers

<table>
<thead>
<tr>
<th>Class No.s</th>
<th>Major covers</th>
<th>Proposed Agglomerative method</th>
<th>Proposed Disaggregative method</th>
<th>Conventional K-means method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Water</td>
<td>488</td>
<td>515</td>
<td>556</td>
</tr>
<tr>
<td>2</td>
<td>Land</td>
<td>403</td>
<td>209</td>
<td>258</td>
</tr>
<tr>
<td>3</td>
<td>Others</td>
<td>31</td>
<td>29</td>
<td>58</td>
</tr>
</tbody>
</table>