Chapter 1

Introduction

1.1 Introduction

The research in the field of Pattern Recognition has got a new dimension with the advancement of today's computer technology, the most fascinating outcome of which is the concept of fifth generation computing system [52], [15], [16]. The ever growing efforts of the scientific and technological community to build an automata that can carry out perception tasks demand the need of enhancement of the knowledge and techniques of Pattern Recognition as much of human behaviour involves its practice.

The application area of Pattern Recognition is quite broad. Automatic speech recognition, Speaker identification, Printed and hand-written character and numeral recognition, X-ray analysis, Signal analysis and various other classification problems are some of the examples. One may visualize the problem of Pattern Recognition to be divided into two parts. One part is concerned with the study of the recognition mechanism as performed by human being and other living organism. The other part is concerned with the development of theory and techniques for designing the machine which can perform the task of recognition automatically. Today's focus is to develop techniques similar to the human mechanism to build machines.

Early research in the field of Pattern Recognition began with the development of mathematical model for recognition process so that perception tasks can be computer implementable. Subsequently, various algorithms have been developed to classify and recognize patterns generated in everyday life. The problem of automatic machine recognition of patterns, in most of the cases basically involves two major tasks, the extraction and selection of significant characterizing features and the design of a suitable classifier. The block diagram of Figure 1.1 represents main concept behind the design of an automatic pattern recognition system. The numerical representation of the pattern con-
stitutes the measurement space. The lower dimensional representation of the measurement space constitutes the feature space and the categorization of the pattern represents the decision space. The accuracy of an Automatic Pattern Recognition system depends jointly on the performance of the feature selector and the classifier. The degradation at any stage affect the efficiency of the other stage resulting in an overall inefficient system. Considerable success has already been achieved in designing classifier. On the other hand designing a feature selector seems to be an extremely difficult task, charged both with theoretical and computational problems. A generalized and effective mathematical theory for feature selection is yet to be developed.

The objective of an automatic pattern recognition system is to categorize some input data into some identifiable classes with the lowest possible probability of error i.e. misclassification. To achieve this goal the system should select those measurements which possess the higher discriminating ability. The purpose of the feature selection stage is manyfold. It's primary justification stems from engineering consideration. The complexity of a classifier and its hardware implementation grows rapidly with the number of dimensions in the pattern space. So it is extremely necessary to reduce the dimensionality of the measurement space. Moreover, dimensionality reduction improves the performance of the classifier, reducing the rate of misrecognition. Furthermore, irrelevant or redundant information carrying measurements may cause detrimental effect in decision process. Thus the importance of feature selection stage cannot be looked down upon. In fact, human recognition act also seems to be basically dependant on some characterizing features of the pattern e.g a few whistled note can evoke the memory of an entire tune or from the blurred image of an object we can identify the object which establishes the human ability of recognising patterns in noisy and incomplete environment. Thus the problem is what the term characterizing means and how to design the criterion for measuring the effectiveness of a particular feature.

Among several approaches proposed so far Statistical Pattern Recognition [13], fukunaga72 has gained the widest acceptance. From the view point of
correct classification those features which will produce lowest error rate of the classifier are considered to be the good features. Most of the pioneer research work that has been done in finding out the suitable criterion for evaluation of the features for the purpose of selection are based on statistical theories like its counterpart in classifier design. Various criteria designed for feature evaluation try to estimate the probability of error associated with a given set of features. But the framework of statistics seems to be inadequate in modelling some of the real life pattern recognition problems.

Another well developed theory has been formulated from the syntactic approach [18] originating from mathematical theory of linguistics. This particular area of pattern recognition has been developed to accommodate non numeric pattern representation and are useful to express structural relationship among patterns.

Among other approaches fuzzy set theoretic approach [56] seems to be well suited in modelling real life problems of pattern recognition where the pattern indeterminacy is due to inherent vagueness or fuzziness in representation rather than randomness in measurement. This approach is also useful in dealing with numeric features as well as non numeric features by assigning numeric membership value to the non numeric feature.

Present day research on Pattern Recognition is motivated from the desire to build computer based machines with perception and cognition capabilities that we ourselves possess. Still now human being definitely does cognitive tasks better than machines, even a three months old baby can somehow recognize the face of her mother. The research on artificial intelligence also have the common motivation i.e. to understand perceptual and cognitive processes in human and to implement similar capabilities in machines. The human ability in pattern recognition seems to depend on learning from experience. To impart human like capability to the machine in performing complex pattern recognition tasks, the research in this area may be guided through the tools of artificial intelligence. The use of domain knowledge in designing automatic feature evaluation system and the design of artificial neural net in learning the characteristic features of a classification problem may provide a link between human pattern recognition and machine pattern recognition. In order to promote research in building modern automatic pattern recognition system proper attention is needed to design efficient feature evaluation techniques as feature selection is considered as the essential preprocessing task of any machine pattern recognition system.

1.2 Objective of the Thesis

In this thesis a study on the problem of Feature Selection in the context of a Pattern Recognition system has been made with a view to provide feature
evaluation measure & consequently feature subset selection techniques from the concept of Fuzzy Set theory. The fuzzy set theory formulated by L.A. Zadeh in 1965 [75] has been applied as an alternative new approach to the problems of Pattern Recognition & Machine learning by several researchers [56]. Statistical tools seem to be inadequate in modelling most of the real life problems in pattern recognition due to the presence of inherent vagueness and ambiguity rather than randomness. These problems are proved to be well handled by the tools of fuzzy set theory.

In the second part of the thesis, the artificial intelligence tools are used to provide modern feature selection techniques in order to impart more human like ability to machines in recognizing patterns generated in every day life. Subsequenly a) a fuzzy artificial neural net model has been proposed to find the importance of features in a feature set via learning mechanism and b) a rule based approach in designing an automatic feature evaluation system using domain knowledge from past experience has been presented.

1.3 Brief Review on Earlier works

The research in the area of automatic Pattern Recognition began during 1950’s gaining momentum from the growing development of the computing ability of Digital Computers. The survey papers of Unger [71], Wee [74], Nagy [50] and Ho & Agarwal [25] represent an account of the earlier works done in this area. The pioneering research on pattern recognition mostly deal with the statistical tools. During 1960’s the main focus of research was on designing classifier, though the problem of feature selection was not totally ignored on those days. Lewis in 1962 proposed mutual information as a measure of feature evaluation. His work was followed by Kamentsky and Liu [30]. Marril & Green [45] proposed divergence as a measure whereas Kailath [29] defined a new distance based measure based on Bhattacharyya distance [4]. The above mentioned papers formed the basis of future developments of various measures based on probabilistic distance and information. The work of Toussaint [69], Matusita [46], Chen [7], Renyi [60] are among the notables. An early work to explore the potential of Karhunen-Loeve expansion in feature selection was done by Watanabe [73] that marked the beginning of another line of research i.e. Mathematical mapping criteria for feature selection. Fu, Min & Li [19] reviewed the contributions made in the first decade of research in the area of feature selection. In 1972 Vilmansen [72] developed some probabilistic dependance measure for feature evaluation. The survey paper of Kittler [39], Chen [6], Chittineni [9] represent an overview of the probabilistic methods developed in seventies. On the other hand the mathematical mapping techniques were further developed by Chien & Fu [8], Fukunaga & Koontz [20] and Kittler &
Young [40]. The survey paper of L. Kanal in 1974 [31] contains exhaustive contributions of statistical and syntactic pattern recognition methodologies developed up to that period.

The features selected on the basis of their individual merit did not guarantee the optimality of the selected feature subset. This fact led to the development of various search techniques for finding optimal feature subset in a given problem. In a large set of features, exhaustive searching requires lots of computational trouble. Thus the mainstream of research on feature subset selection was directed towards developing suboptimal search methods. The invention of Sequential Backward Selection (SBS) in 1963 gave rise to a family of suboptimal stepwise forward and backward methods. All of these methods try to find out the best feature subset by including or removing one feature at a time and the feature to be included or removed being the best or the worst among the existing features at that step. The research in this direction was concluded by introducing the generalization of these algorithm proposed by Kittler in 1978 [13]. Another approach to feature selection based on the concept of dynamic programming was proposed by Chang [5], but this approach burdened by numerous restrictive requirements (e.g., the monotonicity condition and statistical independence of features) has not been heavily pursued by other researchers.

The most promising search technique was proposed by Narendra & Fukunaga in 1977 [51]. Their method known as Branch and Bound Technique guaranteed the selection of an optimal feature subset if the monotonicity condition is satisfied. They originally proposed to use probabilistic separability measures as criterion functions. But their approach has some limitations which was later on partly removed by Foroutan & Sklansky in 1987 [17] by introducing the concept of approximate monotonicity.

Eventually the problem of optimal feature subset selection by nature qualifies as an NP-complete problem. With the development of AI technology the researcher began to try for the solution of the problem with AI approaches. Various search techniques originating from Artificial Intelligence have been explored to examine the potentiality of solving feature subset selection problem. Most recently surprising success has been achieved in solving NP-complete problems in computer science using two classes of techniques simulated annealing and genetic algorithm [38] [24] that simulate natural processes in metallurgy and biology. The overview of research work on methods for automatic feature selection carried out at the Pattern Recognition Project of the University of California, Irvine has been presented by W. Siedlecki & J. Sklansky in their paper in 1988 [68] which covers most of the present trend of the research in this direction.
1.4 Feature Selection: Different Measures

Feature selection is one of the major tasks in any automatic Pattern Recognition system. The main objective of feature selection is to retain the optimum salient characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms for classification can be devised. To facilitate the selection process, the quality of any feature has to be assessed. More important is to design the feature selection criterion for a set of features, because consideration of individual feature does not reveal the redundancy in the input data. The problem of feature selection has two aspects — formulation of a suitable criterion to evaluate the goodness of a feature on one hand and the selection of the optimal subset (based on such criterion) from the possible subsets of the feature space, on the other hand. Thus the general technique of feature selection should follow a two-stage process. In the first stage, a suitable criterion has to be designed to evaluate the quality or goodness of any feature, specially to assess the quality of any selected feature subset. In the second stage a proper search algorithm has to be designed to select the best feature subset among all possible feature subsets. The major mathematical measures so far devised for the estimation of feature quality are mostly statistical in nature and fall under two broad categories:

- Feature selection in the Measurement space.
- Feature selection in the Transformed space.

The main objective of the first category is to reduce the dimensionality of the feature set by discarding features carrying redundant information, the information content of any feature being measured by some chosen criterion, and the features are ordered according to the criterion value to represent the relative effectiveness of the features in a feature set. In contrast, feature selection techniques of the second category utilize all the information of the pattern vector and map higher-dimensional pattern vector to a lower-dimensional one. A broad and complete overview of the techniques are available in the book by Kittler [39]. In the following two subsections, the common statistical measures for feature evaluation are briefly presented. The next subsection represents a brief overview of the feature selection methods in transformed space.

1.4.1 Feature Selection in the Measurement space

Error Probability

Pattern classification is a decision-making process in which an input pattern in an $n$-dimensional feature space, defined by $f = [f_1, f_2, \ldots, f_n]$, is assigned
to one of a number of possible classes \( W_i \), where \( i = 1, 2, \ldots, m \), depending on the value of the feature vector. Let the a priori probability of occurrence of a class \( w_i \) be \( P(w_i) \), and let the multivariate class-conditional probability density function of class \( w_i \) be \( p(f | w_i) \). The mixture density function \( p(f) \) is given by

\[
p(f) = \sum_{i=1}^{m} p(f | w_i)P(w_i)
\]

According to Bayes' rule [14], the a posteriori probability for the \( i \)th class is

\[
P(w_i | f) = p(f | w_i)P(w_i)p(f)
\]

and according to Bayes' decision procedure, an unknown input pattern is assigned to class \( w_i \) if

\[
P(w_i | f) \geq P(w_j | f) \text{ for all } i, j = 1, 2, \ldots, m, i \neq j
\]

The objective of any pattern recognition scientist is to design a system which will classify unknown patterns with the lowest possible probability of misrecognition. Being a subsystem the objective for designing the feature selection criterion should be such that the error probability \( p_e \) (the probability of misrecognition) is minimised. In an \( n \)-dimensional feature space \( f \), the error probability \( p_e \) is given by

\[
p_e = \int [1 - \max_i P(w_i | f)]p(f)df
\]

Although the error probability is, from the theoretical viewpoint, the ideal measure for designing the feature selection criterion, it is not easy to evaluate from the practical computational point of view. Therefore a number of alternative criteria for feature evaluation have been suggested so far. Most of the indirect measures have been developed based on the concept of distance, separability, overlap or dependence between the probability distributions characterising the pattern classes.

**Interclass distance**

The discriminatory power of any feature is associated with the concept of class separability. The interclass distance between two classes \( w_1 \) and \( w_2 \) is the simplest concept of class separability and can easily be used to assess the discriminatory potential of a feature in pattern representation. Intuitively, the more distant the classes from each other, the better the chances of successful recognition of class membership of the patterns. This measure is not defined explicitly by class-conditional probability density functions. Its estimate based on elements of the training set can be computed directly without prior determination of the probabilistic structure of the classes. Separability measures
based on this concept thus cannot serve as true indicators of the mutual overlap of the classes, as the detailed information of the classes are not used here, though it is very simple to evaluate.

The measures derived from the concept of class separability that are normally used in practice are divided into the following three groups:

**Measures derived from the Probabilistic distance**

To obtain a realistic picture of mutual class overlap, measures have been developed based on the concept of the measure of distance between the probability density functions characterising the pattern classes. These measures are termed as *probabilistic distance measures*. They do not bear any exact relationship to $p_e$ and hence upper and lower bounds expressed in terms of these measures, have been derived to provide an indication of the accuracy of the estimate of $p_e$. The error probability $p_e$ in the two-class case is easily shown as, for $m=2$,

$$p_e = 0.5[1 - \int |p(f|w_1)P(w_1) - p(f|w_2)P(w_2)|df]. \quad (1.5)$$

Now $p_e$ is maximum when the integrand is zero, i.e., the density functions are completely overlapping, and $p_e$ is zero when the functions do not overlap. The integral in equation 1.5 can be considered to quantify the 'probabilistic distance' between the two density functions. The greater the distance, the smaller the error and vice versa. Any other measure of 'distance' between the two density functions such as,

$$J = \int f[p(f|w_i), P(w_i), i=1,2]df, \quad (1.6)$$

satisfying $J \geq 0$ can be used as a feature evaluation criterion. $J = 0$ when $p(f|w_i), i=1,2$ are overlapping and $J$ is maximum when $p(f|w_i), i=1,2$ are nonoverlapping.

The most commonly used probabilistic distance measure are given below:

1. Bhattacharya Distance : The Bhattacharya distance function [4] is defined as follows:

$$J_B = -\ln \int [p(f|w_1)p(f|w_2)]^{1/2}df. \quad (1.7)$$

For Multivariate Gaussian distributions, that is, where

$$p(f|w_i) = N(\mu_i, \Sigma_i), i=1,2,\ldots, \quad (1.8)$$

$J_B$ becomes

$$J_B = 1/8(\mu_1 - \mu_2)^T \left( \begin{array}{cc} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_2 \end{array} \right)^{-1}(\mu_1 - \mu_2) + 1/2 \ln \left\{ \frac{\det(\Sigma)}{\sqrt{\det(\Sigma_1)det(\Sigma_2)}} \right\}. \quad (1.9)$$
$\mu_i$ and $\Sigma_i$ represent mean vector and covariance matrix for the probability distribution of the feature for $i$th class respectively and $I$ represents the identity matrix.

2. The Jeffreys-Matsusita Distance measure: The Jeffreys-Matsushita Distance measure \cite{28} $J_M$ is defined to be

$$J_M = \int \left( \sqrt{p(f|w_1)} - \sqrt{p(f|w_2)} \right)^2 df. \quad (1.10)$$

It is to be noted that the Jeffreys-Matsusita distance and Bhattacharya distance are two different variations of the same measure and the relationship between them is given by

$$J_M = \{2[1 - \exp(1 - J_B)]\}^{1/2}. \quad (1.11)$$

3. Divergence function: The divergence function was first introduced by Jeffreys \cite{28} and is defined by

$$J_D = \int [p(f|w_1) - p(f|w_2)] \ln \frac{p(f|w_1)}{p(f|w_2)} df. \quad (1.12)$$

Later it was developed by Toussaint \cite{70} and for the multivariate Gaussian distribution, the divergence function becomes

$$J_D = \frac{1}{2} \text{tr}(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I) + \frac{1}{2} \text{tr}[(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T]. \quad (1.13)$$

4. Mahalanobis Distance: For two multivariate distributions with common dispersion matrix, the Mahalanobis distance \cite{44} is given as

$$D^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2). \quad (1.14)$$

This distance function is very easy to compute though it is difficult to make theoretical assessment of the accuracy in a distribution-free case.

5. Kolmogorov variational distance: The Kolmogorov variational distance \cite{13} is defined to be

$$J_K = \frac{1}{2} \mathbb{E}\{|(P(w_1|f) - P(w_2|f)|)\}. \quad (1.15)$$

From this measure, the probability of error $p_e$ can be determined directly by the equation

$$p_e = \frac{1}{2} - J_K, \quad (1.16)$$

but the computational difficulty associated with this measure is the same as that for $p_e$. 

16
There are several other measures, developed from time to time by the re­searchers, a complete discussion on their relative merits and demerits in eval­uating feature quality being available in [9]. In this present study compar­ison has been made with the above listed statistical methods. All of those probabilistic distance measures become important when the class conditional probability distributions are parametric and in particular normal (Gaussian) In that case, integrals in their expressions can be evaluated analytically, thus avoiding a tedious numerical integration.

### Measures derived from Probabilistic dependence

The degree of dependence between the feature vector and the class membership can be used as a measure of the effectiveness of any feature in distinguishing different classes. The degree of dependence between the variables $f$ and $w_i$ can be measured by the distance between the conditional density $p(f|w_i)$ and the mixture density $p(f)$. Any of the probabilistic distance measure can be used for evaluating the probabilistic dependence between $f$ and $w_i$ simply by replacing one of the class-conditional density functions with the mixture density. Some of the measures are given below. By computing the weighted average of the class-conditional distance the overall dependancy can be used which will indicate the effectiveness of any feature in a multiclass environment. Some of the measures are given below:

1. Bhattacharyya:

\[
I_B = \sum_{i=1}^{c} P(w_i) \left[ - \ln \int \sqrt{p(f|w_i)p(f)} df \right]
\] (1.17)

2. Matushita Distance:

\[
I_M = \sum_{i=1}^{c} P(w_i) \left\{ \int [\sqrt{p(f|w_i)} - \sqrt{p(f)}]^2 df \right\}^{0.5}
\] (1.18)

3. Patric-Fisher Distance

\[
I_P = \sum_{i=1}^{c} P(w_i) \left\{ \int [p(f|w_i) - p(f)]^2 df \right\}^{0.5}
\] (1.19)

All the above measures correspond to a multiclass environment where the total number of classes is $m$.

### Measure derived from Information Theoritic Approach

From the view point of Information Theory ( originated by famous scientist Shannon [67] ), information gain is analogous to reduction in uncertainty which
in turn is quantified by entropy. From the posteriori probabilities $P(w_i|f)$, one can gain information about the dependance of $f$ on $w_i$. Thus the expected value of the generalized entropy function can be used as a measure of feature quality, considering that, the smaller the uncertainty, the better the feature vector. The average generalized entropy of degree $\alpha$ is defined as

$$J^\alpha_E = \int (2^{1-\alpha} - 1)^{-1} \sum_{i=1}^{N} P^\alpha(w_i|f) - 1 |p(f) df,$$  \hspace{1cm} (1.20)

where $\alpha$ is a real, positive parameter. For $\alpha = 1$, the measure becomes

$$J_s = - \int \sum_{i=1}^{N} P(w_i|f) \log_2 P(w_i|f) df,$$  \hspace{1cm} (1.21)

which is the separability measure based on the well-known Shannon entropy. Optimal feature set can be obtained by minimizing the above criterion.

### 1.4.2 Feature selection in the transformed space

The main idea behind this group of techniques is to map a higher-dimensional data into a lower-dimensional one, so that further processing of the data for classification requires less computational complexity. Apart from that, in some cases, to separate the data collection stage from the classification stage of a pattern recognition system, it may be highly desirable to compress the observed information into a lower-dimensional space to facilitate its transmission or storage. The mathematical mapping techniques are computationally simpler than the probabilistic criteria and produce satisfactory results in practical situations, but unlike the probabilistic criteria they do not have any relationship with the error probability. In the present thesis, the mapping techniques have not been studied for comparison. For the sake of completeness only an introduction to the main methods of the mapping techniques are included here. The mapping methods are mainly divided into two classes: Linear transformation & Non linear transformation. Among linear mapping methods the main techniques are:

**Diagonal Transformation**

One technique for increasing the similarity among the points in a class is the minimization of a metric between the two points defining the class. Such an intraclass measure has been defined by Sebestyn [66] and is based on a Euclidean square metric. In order to minimize such an intraclass measure of a class of points in a multidimensional space, each dimension in the pattern space will be given a variable weightage. This is equivalent to multiplication of the pattern space by a diagonal matrix & hence the name.
Rotational Transformation

This technique is based on an orthogonal transformation equivalent to a rotation of the original pattern space to a new set of coordinate vectors which are also orthogonal and which hopefully provide a domain for easier feature identification, and is more amenable to simpler classification algorithm. The transformation is chosen in a way which approximates the original prototypes with the fewest dimension in a mean square error sense.

Nonlinear Transformation

One of the major disadvantages with linear transformations is that they often do not separate prototypes of classes that are clearly separable but which do not separate due to rotations, scaling or other linear operations. The nonlinear transformation transform the pattern space into a feature space so that nonlinearly separable classes become linearly separable. An example of simple nonlinear transformation is given by Sammon [65].

1.5 Feature Subset Selection: Search Algorithms

The ultimate objective of the feature selection stage in an automatic pattern recognition system is to find out the best combination of $M$ features out of possible $N$ features in order to reduce the dimensionality as well as to remove irrelevant informations. The discriminatory power of any feature set has to be assessed by some suitable evaluation criterion and a search procedure has to be adopted to find out the best subset of features among all possible subsets. The easiest way to achieve this is to select $m$ individually best features out of $n$ possible features according to certain criterion. More specifically, if features are ranked according to their importance by some evaluation criterion and if the ranking be $f_1 > f_2 > f_3 \cdots > f_N$ (from higher to lower importance). Then the set $F = (f_1, f_2, \ldots, f_M), M \leq N$ may be selected as the best subset of $M$ features out of $N$ features. But even if, the features are independent this method does not guarantee the optimality of the selected subset. When the number of features is large, exhaustive search among all possible subsets requires excessive computations. So various search procedures are designed to obtain suboptimal or optimal solution.
1.5.1 Suboptimal Search Algorithms

In many situations the determination of the optimal feature set may not be computationally feasible. In such situations suboptimal search methods are used. A brief survey of the simple suboptimal search techniques are presented below.

Sequential Forward Selection (SFS)

SFS is a simple bottom up search procedure where one measurement or feature at a time is added to the current feature set [13]. Starting from an empty set, the first feature selected is the individually best feature. At each subsequent stage the next feature is selected from the remaining available features so that in combination with the features already selected the new set yields the best value of the criterion function.

Let us suppose that $M$ features to be selected out of $N$ features of the feature set $f = (f_1, f_2, \ldots, f_N)$. The algorithm is as follows:

Step 1. Rank all the features according to the selection criterion. Let $f_k$ represents the best feature i.e $J(f_k) = \max_i J(f_i), i = 1, 2, \ldots, N$.

Step 2. Compute the selection criterion for the feature set $(f_k + f_i)$ where $f_i \in f, i = 1, 2, \ldots, n, i \neq k$ and rank them. Let $J(f_k + f_i) = \max_j J(f_k \cup f_j), f_j \in f$ where $j = 1, 2, \ldots, N, \text{ and } \neq k$. $(f_k + f_i)$ is selected as the best set of two features at this step.

Step 3. Step 2 is repeated to select one feature at each step until a feature set of $M$ features has been selected.

This algorithm is definitely better than selecting the best $M$ features in the individual ranking. The main drawback of SFS is that, once a feature is included in the feature set, the method has no mechanism for deleting it from the feature set even if at a later stage it becomes superfluous due to addition of other features.

Generalized Sequential Forward selection GSFS($r$)

GSFS($r$) algorithm is basically a SFS method but here, instead of adding one feature at a time, $r$ features are added to the current feature set at each stage of the algorithm. The algorithm is as follows:

Step 1. Select $k$ feature set from $N$ features by evaluating the criteria function for all feature sets of $k$ features.

Step 2. Generate all possible sets of $(k + r)$ features by adding the selected $k$ feature set to combinations of $r$ features from the set of available features $(n - k)$. Then select the feature set of $(k + r)$ features in such a way that the set maximizes the selection criterion among all other combinations.
Step 3. Step 2 is repeated until the required number of features are selected. GSFS is likely to be more reliable than SFS, but of course, more costly to implement, for at each stage \((n - k) C_r\) feature sets have to be inspected to find feature set of \((k + r)\) elements. Here also the feature once included cannot be deleted at a later stage.

Sequential Backward Selection (SBS) and Generalized Sequential Backward Selection (GSBS)

SBS is the top-down counterpart of the SFS method. Starting from the complete set of features, one feature is eliminated a time.

At each stage the measurement selected for elimination is the one that results in the lowest decrease in the value of the criterion function. By analogy, the GSBS algorithm is the same as the GSFS with the exception that \(r\) features are discarded at each step instead of being added.

From the point of view of computational complexity, SFS is simpler than SBS since it requires that the criterion function be evaluated at most in \(M\)-dimensional space. In contrast, in SBS the criterion function must be computed in spaces of the dimensionality ranging from \(M\) to \(N\). The source of suboptimality is also the same in both the cases, as once discarded, SBS does not allow any feature for reentry. However, in SBS, at each stage the information loss can be assessed.

Plus \(l\) - Take Away \(r\) Algorithm

The nesting of feature sets, which may quickly result in suboptimality of sequential selection algorithms, can be partially, overcome by alternating the process of addition and removal of features from the feature set. After adding \(l\) measurements to the current feature set, \(r\) features are then removed. Thus the net change in feature set size is equivalent to \(l - r\). This process is continued until the feature set reaches the required size. If \(l > r\), the feature selection procedure operates bottom up, and for \(l < r\), it operates top down.

The algorithm is as follows:-

for \(l > r\), Step 1. let the current feature set be \(F_k\) containing \(k\) features.
Step 2. Apply SFS algorithm \(\ell\) times to generate feature set of \((k + \ell)\) features.
Step 3. Apply SBS algorithm \(r\) times to generate feature set of \((k + \ell - r)\) features.
Step 4. Repeat step 2 and step 3 until the required number of features \(M\) in the feature set is achieved.
for \(l < r\),
The same as above with step 2 and step 3 interchanged.
When $r = 0$, the algorithm becomes SFS algorithm and when $\ell = 0$, the algorithm becomes SBS algorithm.

**Generalized Plus $\ell$ - Take Away $r$ Algorithm**

The only difference between this algorithm and the former one is that here instead of employing SFS and SBS algorithms for number of times, GSFS and GSBS algorithms are employed in a number of substeps.

The algorithm is as follows:-

1. Let the current feature set be $f_k$ containing $k$ features.
2. Apply GSFS algorithm to add $\ell$, features at a time and the process is repeated $k$ times where $\ell \times k = \ell$.
3. Apply GSBS algorithm to remove $r_j$ features at a time and the process is repeated $p$ times where $r_j \times p = r$.
4. Step 2 and step 3 are repeated until the feature set contains the required number of features.

These algorithms remove the main drawback of SFS and SBS algorithms to certain extent.

**Max - Min Feature Selection**

This algorithm is based on the feature selection on the individual and pairwise merit of measurement. The algorithm is as follows:-

1. Let $F_k$ be the current feature set, containing $k$ features.
2. Select as the $(k + 1)$st feature that measurement $f_j \in (F_k)$ which satisfies $\delta J(f_{k+1}, f_j) = \max_i \min_j \delta J(f_j, f_i)$, $f_i \in F_k$.
3. Step 2 is repeated until required number of features in the feature set is selected.

The heuristic basis is as follows. After selecting $k$ features the merit of the remaining features can be assessed by measuring the amount of new information they would contribute if added to the feature set.

**1.5.2 Optimal Search Algorithm**

**Branch and Bound Algorithm**

The only method which guarantees the optimality of a selected subset without explicit evaluation of all the possible feature subsets was proposed by Narendra & Fukunaga [51]. The algorithm known as Branch and Bound algorithm is applicable under the assumption that the feature selection criterion satisfies the monotonicity property. Denoting by $F_k$ the feature set containing $k$ features,
the monotonicity property implies that the criterion function $J$ used to evaluate feature subsets change monotonically over a sequence of nested feature subsets

$$(F_1 \subseteq F_2 \subseteq F_3 \ldots \subseteq F_k \Rightarrow J(F_1) \leq J(F_2) \leq J(F_3) \leq \ldots \leq J(F_k) \quad (1.22)$$

(Considering the criterion function increases with the inclusion of features in the set). This condition merely means that a subset of features should not be better than any larger set that contains the subset.

The Branch and Bound algorithm as applied by Narendra & Fukunaga successively generates a portions of the solution tree, by rejecting one feature at a time from the complete set until the final stage at which the subset contains the required number of features. The algorithm computes the criterion value at each node. Whenever the criterion value for any node is less than the pre-set threshold $B$, all nodes, successors of that node will have criterion value less than $B$ and that portion of the tree will not be explored. Thus the algorithm backtracks and computes the criterion value for the hitherto unexplored nodes until the solution space is exhausted. By suitable generation of the solution tree and optimal organisation of the search procedure the algorithm can be made more effective by early backtracking and rejecting partial sequences containing more nodes without affecting the optimality of the selected subset.

To find out an optimal subset of $M$ features from a set of $N$ features certain strategy should be adopted for an efficient search. First of all the generation of the solution tree should be systematic to ensure easy programmability. Secondly as the section of the tree branching from a node having low value of criterion function is more likely to be rejected than the branches originating from a node having larger criterion value, it is logical to attach larger number of successors to the node having lower value of the criterion function. In other words, the tree should be asymmetrical. Thirdly the tree must have $n - m$ levels such that the terminal nodes correspond to the feature set containing $m$ features.

The tree is actually constructed from the least dense part of the tree to the part with most branches (from right to left as shown in Fig 1.2). First the single rightmost branch is generated and the magnitude of the separability measure at the terminal node is taken as the current threshold. Then the algorithm backtracks to the nearest branching node and generates the next rightmost branch of the tree. If, at any node the criterion function has the value less than the threshold, the search of the partial tree originating from that node is abandoned and the algorithm backtracks to the nearest branching point in next lower level. The threshold is updated and the criterion value of that node is taken as the current threshold and the process of generation of the tree is continued until the whole tree is generated.

Formally the algorithm is as follows:-
Figure 1.2: Generation of Solution tree in Branch and Bound technique
Let \((f_1, f_2, \ldots, f_M)\) be the \(M = N - M\) features to be discarded to obtain an \(M\) feature subset and \(J_M(f_1, f_2, \ldots, f_M)\) represents the feature selection criterion of the feature set containing the features other than \((f_1, f_2, \ldots, f_M)\).

List\(i\): An ordered list of the features to be discarded at level \(i\).

Pointer\(i\): The pointer to the element of List\(i\) being currently considered.

Successor\(i, k\): The number of successor that the \(k\)th element in the List\(i\) can have.

Avail: A list of available feature values that List\(i\) can assume.

Step 0. (initialization) Avail = \(\{1, 2, \ldots, N\}\), \(i = 1\), List\(0\) = \{0\}, Successor\(0, 1\) = \(M + 1\), Pointer\(0\) = 1, set bound = \(B\).

Step 1. (initialize List\(i\)) Set Node = Pointer\(i - 1\).

Compute \(J_i(f_1, f_2, \ldots, f_{i-1}, f_k)\) for all \(k\) in Avail. Rank the features in Avail in increasing order of the criterion function and store the smallest \(p\) of these in List\(i\) in the increasing order (with the first element in List\(i\) being the feature in Avail yielding the smallest value of the criterion function), where \(p = \text{Successor}(i - 1, \text{Node})\). Set Successor\(i, j\) = \(p - j + 1\), for \(j = 1, 2, \ldots, p\). Remove List\(i\) from Avail.

Step 2. (select new node) If List\(i\) is empty, go to Step 4. Otherwise, set \(f_i = k\) where \(k\) is the last element in List\(i\). Set Pointer\(i\) = \(j\) where \(j\) is the current number of elements in List\(i\). Delete \(k\) from List\(i\).

Step 3. (check bound) If \(J_i(f_1, f_2, \ldots, f_i)\) is less than the bound \(B\), return \(f_i\) to Avail. Go to Step 4. If level \(i = M\), go to Step 5. Otherwise, set \(i = i + 1\) and go to Step 1.

Step 4. (backtrack) Set \(i = i - 1\). If \(i = 0\), go to Step 6. Otherwise return \(f_i\) to Avail and go to Step 2.

Step 5. (final level, update bound) Set \(B = J_M(f_1, f_2, \ldots, f_M)\) and save the corresponding feature set as the candidate set for the optimal feature subset.

Step 6. (termination) Terminate the algorithm and declare the final candidate set used for the last updating of the bound as the optimal feature subset.

Narendra & Fukunaga originally proposed to use probabilistic separability measures as criterion function. Recursive equations were derived (for ease of computation) for the parametric criterion function such as the Bhattacharyya distance and the Divergence that are monotonic with respect to the feature subset.

**Approximate Monotonicity**

In Branch and Bound technique the optimality of the selected subset is constrained by the requirement of monotonic criterion function. But this is an major obstacle in many practical situations where the sample size is small and no a priori information about the data is available. The estimation of the probability density function based on the sample of finite size introduces some error...
which can turn a monotonic criterion function into a nonmonotonic one. The use of error probability (error rate) as the evaluation criterion may not remove the difficulties as due to same origin, it may lose monotonicity. Foroutan & Sklansky [17] introduced the concept of approximate monotonicity and showed that locally trained piecewise linear classifier yields approximately monotonic resubstitution error rate with respect to the number of features. Their technique can be successfully used for selecting near optimum feature set for the classification of non-gaussian data and is really an improvement over Narendra & Fukunaga's technique.

1.6 Modern Approaches to Feature Selection

In the previous section Branch and Bound methods with the concept of approximate monotonicity has been discussed. For a large scale problem (number of features greater than 20), the number of nodes to be examined before finishing with a optimal subset may be very large as there is no guarantee of stopping the algorithm within a predefined time limit. In this situation Branch and Bound method needs to be modified to design variable feasibility threshold so that the algorithm would be very aggressive at the beginning of deletion of nodes and less aggressive at its end. The capability of other search algorithms in optimization problem originating from artificial intelligence techniques have also been explored in feature selection research to find out optimal feature subset in a given problem.

Simulated Annealing

Simulated Annealing [38] has been developed to solve NP-hard combinatorial optimization problem. In simulated annealing the analogy is drawn between the solution space of a combinatorial optimization problem and the atoms or molecules of a system tending to cool to an optimal equilibrium. Annealing, in its original formulation is concerned with the behaviour of thermal systems in low temperature equilibrium. To analyze the behaviour of a collection of atoms in equilibrium at a given temperature Metropolis et. al [48] simulated the behaviour of thermal system through the use of probabilistic model. Simulated annealing algorithm transforms an optimization problem, in a problem specific manner into an annealing problem. In each case certain objective function is to be minimized which corresponds to the internal energy in case of thermal system and near optimal solution corresponds to the stable low temperature equilibrium state. This algorithm can be applied to the problem of feature subset selection.
Genetic Algorithm

Like the Simulated Annealing approach, the Genetic Algorithm [11], [23] is another Artificial Intelligence approach for examining a state space to obtain a good solution. A genetic algorithm is a parallel test and go technique, in which a predefined number of possible solutions is modified, tested and stored at the same time. The solution, which is usually represented by a finite sequence of numbers (i.e. a point in the state space), is called a chromosome. The algorithm manipulates a set of chromosomes, the population, in a manner resembling the mechanism of natural evolution. The chromosomes are allowed to mate or crossover, and mutate. The mating of two chromosomes produces an offspring chromosome which is a synthesis of its parents. A mutation of a chromosome produces a near identical copy with some components of the chromosome altered. An exponential growth of the system of chromosomes is achieved by introducing the score function, which is basically the function to be optimized, and by limiting the size of the population. In other words, each chromosome is evaluated and only a predefined number of the best ones survive to the next cycle of reproduction. The population of chromosomes is evolved through a number of such cycles called generations. Despite the limited size of the population, it is capable of fast adaptation, resulting in rapid optimization of the score function.

As in the simulated annealing algorithm the proper choice of operators and evaluation functions is the key to creation of an application oriented version of the genetic algorithm. This algorithm can be used for feature selection problem.

1.7 Outline Of The Thesis

The present chapter begins with the introduction of the problem and its importance to the context. In the next section the objective of the present thesis has been stated. The chapter then continues with a brief review of the earlier works done and a brief sketch of the contents of the following chapters. The following section contains the commonly used methodologies for the evaluation of effectiveness of features in a pattern recognition problem. No attempt has been made here to present an exhaustive catalogue of all the existing techniques, but most of the popular statistical measures for feature evaluation designed so far have been included and their relative merits and demerits have been discussed. A short account of the search techniques have been given with a special emphasis to the most useful one i.e. the Branch and Bound technique which has been explained in greater detail. The chapter concludes with a brief reference to the modern approaches from the corner of Artificial Intelligence techniques.
to the problem of feature selection.

The second chapter begins with the introduction of the concept of the Fuzzy set theory and its application to the present problem. The basic operations on fuzzy set has been stated and the measures of fuzziness of a fuzzy set has been given. The Index of Fuzziness (\( \nu_A \)) and the entropy (\( H_A \)) of a fuzzy set \( A \) are the terms defined by Deluca & Termini to quantify the amount of fuzziness associated with the set. The concept of intraset and interset ambiguity has been introduced and their evaluation in terms of measures of fuzziness has been presented. In the following section the proposed measure for evaluating feature quality has been designed in terms of intra and interset ambiguities. In the final section the algorithm has been implemented on three different data sets and the features for a particular data set are ordered according to their effectiveness using the proposed measure. The results obtained have been discussed in the conclusion, explaining the applicability of the proposed measure.

In the next chapter some of the common statistical measures have been selected for finding out the feature ranking according to their effectiveness in discriminating between any two pattern classes from a given set of features. The measures are then applied to three data sets (the same ones are used in the previous chapter) and the features in a given set are ranked according to the different measures. These rankings are then compared with the ranking obtained via the proposed measure in a particular feature set in a given problem. The conformity between two rankings have been tested via statistical rank correlation analysis and recognition experiment. At the conclusion the advantages of the proposed technique over statistical techniques are pointed out.

In the fourth chapter the problem of feature subset selection have been addressed. The fuzzy set theoretic criterion function proposed for feature ranking in Chapter 2. has been extended in multidimension using multidimensional membership function to find the effectiveness of a particular feature set containing a number of features. Feature sets are ranked using this criterion function. In the next step this criterion function in conjunction with the branch and bound technique is applied to find out the best feature subset of \( M \) features from a set of \( N \) features and to show the effectiveness of the proposed criterion function over statistical ones for using with branch and bound technique. The algorithm has been implemented with the vowel data set with extended feature set for vowel and speaker recognition problems. Though the feature set contains small number of features it establishes the effectiveness of the algorithm and can be extended to use with any number of features.

The next chapter introduces knowledge based approaches to the problem of feature selection. A guideline for designing automatic feature selection system with the help of AI and Neurocomuting methods has been developed
1.8 Conclusion

In this Chapter a short account of the different approaches to the problem of feature selection developed from the early period to the current time has been presented. In the first part attention has been given to discuss about various measures developed from statistical theory to assess the discriminatory power of any feature in a given feature set. The features are to be ordered according to those criterion function and the ordering is important when applied to sequential classifier for decision purpose. In the second part the feature selection problem is viewed as a search for the best subset of features among a number of feasible subsets. Several search techniques with optimal and non optimal solutions are discussed. Two approaches (Simulated annealing and Genetic Algorithm) have been named as currently these two artificial intelligence approaches have been successfully applied to other combinatorial optimization problem. They can be potentially used to the feature selection problem.

In the following Chapter fuzzy set theoretic concept has been used to formulate feature evaluation criterion to cope with real life problems in which the features may be numeric as well as non numeric.