CHAPTER 2

Pattern Classification Techniques: An Overview

As mentioned in the introductory chapter, we are mainly concerned with the problem of pattern classification. Pattern classification is essentially viewed as partitioning of the feature space or mappings from feature space to the decision space. The various techniques for pattern classification may be broadly classified as -- Deterministic techniques, Statistical techniques, techniques based on fuzzy sets and Linguistic techniques. For the first three techniques mentioned above, the pattern classes are generally considered to be sets of points in the M-dimensional feature space, i.e., a pattern \( X \) would be viewed as an M-dimensional vector -

\[
X = \begin{bmatrix} x_1, x_2, \ldots, x_M \end{bmatrix},
\]

where \( x_j \) is the feature of the pattern in the \( j^{th} \) axis of measurement. However, it would be observed in the sequel that the Linguistic techniques do not assume the patterns to be points in the M-dimensional feature space, rather they are viewed as strings of pattern primitives.

The proposed hybrid classifier uses a technique based on
fuzzy sets and a linguistic technique during two different phases of classification. As the notion of fuzzy sets and concept of pattern classification using fuzzy set theory is relatively recent, a complete section is devoted to the concepts of fuzzy sets and pattern classification using fuzzy sets. The volume of literature on linguistic techniques in pattern classification is quite enormous, so we review only the basic concepts and important results in this field. In section 2.1 we briefly present the deterministic and statistical techniques indicating the most important results and their limitations. This section is included primarily for the sake of completeness, as well as with the objective of providing motivation for using fuzzy set based techniques.

2.1 Deterministic & Statistical Classification Techniques.

In the following paragraphs we will briefly review some of the important results in Deterministic and Statistical techniques used for pattern classification. The Deterministic techniques are based on the assumption that the feature vectors under consideration are deterministic in nature. Where as the Statistical techniques have been developed for classification of patterns when the representative feature vectors are random variables. For simplicity of notation, we will present the classification techniques for a two class problem. The pattern classes under consideration will be denoted as $C^1$ and $C^2$. 
The generalization of these techniques for multiclass problems readily follows.

**Deterministic Classification Techniques:**

In general, the deterministic techniques for pattern classification are formulated in terms of **discriminant functions** [Nilsson (1965), Meisel (1972)]. The discriminant functions \( D_1(X) \) and \( D_2(X) \) associated respectively with the pattern classes \( C^1 \) and \( C^2 \) are such that — if the pattern represented by the feature vector \( X \) is in class \( C^1 \) (denoted by \( X \in C^1 \)) then \( D_1(X) > D_2(X) \) … (2.1)

Various forms of discriminant functions have been reported in the literature. The simplest class of discriminant functions are the linear discriminant functions. A linear discriminant function is expressed as a linear combination of the feature measurements \( x_1, x_2, \ldots, x_M \).

The linear discriminant functions may be expressed as

\[
D_i(X) = \sum_{k=1}^{M} w_{ik} x_k + w_{i,M+1} \quad \text{for } i = 1, 2, \ldots \quad (2.2)
\]

where the parameters \( w_{ik} \)'s are often termed as weighting parameters.

The decision boundary between the region in feature space associated with the classes \( C^1 \) and \( C^2 \) may be expressed as \( D_1(X) - D_2(X) = 0 \), which is essentially the equation of a hyperplane in the feature space.

When there exists hyperplane separating the classes under consideration, then the classes \( C^1 \) and \( C^2 \) are said to be linearly separable.
i.e., the feature vectors corresponding to \( C^1 \) lie on one side of the hyperplane and the feature vectors for class \( C^2 \) lie on the other side. If it is possible to obtain the equation of the hyperplane, then it is possible to achieve correct classification.

In practice, when we assume the discriminant function to be linear, the values of the weighting parameters are not known apriori. Under this circumstance, the weighting parameters are estimated by presentation of sample patterns of known classification. These patterns of known classification are termed as training patterns. The method of estimation of unknown parameters using the patterns of known classification is known as supervised learning of the classifier (learning with a teacher). The performance of the classifier improves as more and more training patterns from both the classes are presented. If the two classes under consideration are linearly separable then there exists a solution weighting vector, which may be estimated by the presentation of a finite sequence of training patterns from both the classes [Novi Koff (1963)]. An exhaustive treatment on training of linear classifiers may be found in Nilsson (1965), Ho and Agrawala (1968).

If the pattern classes under consideration are not linearly separable, the use of piece wise linear or other nonlinear discriminant
functions have been reported in the literature [Hoffman and Moe (1969), Duda and Fossum (1966)]. In these situations, the results from the theory of nonlinear programming are applied in estimation of a suitable nonlinear discriminant function. However, in most cases, it is difficult to establish the convergence of these estimation algorithms. The algorithms may only converge for a highly restricted class of admissible functions. Moreover, the computational complexity of the algorithms for estimation of nonlinear functions is prohibitive.

**Statistical Classification Techniques:**

The Statistical classification techniques, in general, deals with a further generalization of the discriminant function approach, where the unknown parameters in the discriminant functions are estimated from the statistical information regarding the pattern classes. The various classification techniques to be discussed require or assume different type of statistical information regarding the pattern classes under consideration. In this context, the problem of pattern classification may be viewed as the estimation of the decision function (essentially the discriminant function) \( f(X) \) from the available information, such that

\[
\begin{align*}
\{ f(X) &> 0 \text{ for } X \in C^2 \\
\{ f(X) &< 0 \text{ for } X \in C^1 \\
\end{align*}
\]

... (2.3)
When the conditional density function \( p(X|C^i) \), \( i=1, 2, \ldots \), for both the classes are completely known, then the problem reduces to that of simple hypothesis testing in Statistics, and the decision function \( f(X) \) is usually selected as

\[
f(X) = L(X) - \eta, \quad \ldots \quad (2.4)
\]

where \( L(X) \) is the likelihood ratio defined as

\[
L(X) = \frac{p(X|C^2)}{p(X|C^1)} \quad \ldots \quad (2.5)
\]

and \( \eta \) is a threshold value. In the statistical literature various choice of \( \eta \) have been suggested depending on Neyman–Pearson Criterion, Baye’s Criterion or the Minimax Criterion [Hogg & Craig (1970)]. If \( p(X|C^i) \) is a Gaussian distribution function, the likelihood ratio can be explicitly written in terms of the means and covariances. If the covariances for both the classes are assumed to be equal, the decision function \( f(X) \) reduces to a linear discriminant function, where the weighting parameter may be uniquely determined by the mean and covariance of the Gaussian distribution function and the value of \( \eta \).

The choice of the best linear function, based on the given set of samples for the Gaussian case is dealt in Anderson and Bahadur (1962).

As pointed out by Fu (1968), it is often necessary to estimate the decision function when the patterns \( X_k \) are received
sequentially. The main tool used in such a situation is the sequential probability ratio test (SPRT) developed by Wald (1947). Computationally the main problem in the use of SPRT is the recursive evaluation of the likelihood function. When the patterns $X_k$ belong to a Gaussian-Markov sequence, then it may be readily shown that $-\frac{1}{n} \left[ L_{k+1} (X) \right]$ is a linear combination of $\frac{1}{n} \left[ L_k (X) \right]$ and a term $\frac{1}{n} (X_{k+1})$, where

$$L_k (X) \triangleq L(X_1, \ldots, X_k) = \frac{p(X, \ldots, X_k | C)}{p(C)} \quad \ldots (2.6)$$

Fu and his associates have studied various aspects of the sequential method as applied to pattern recognition \cite{Fu 1968}. Fu and Chien (1966) have further generalized the SPRT when the threshold value $\gamma$ is assumed to be a function of time and have discussed the method for recursive estimation of $\gamma$ from the sequential sample presentation.

The situation which is often encountered in pattern classification problems is that of the estimation of the decision function when only the functional form of $p(X | C, \emptyset)$ is given with the unknown parameters $\emptyset$. For example, it may be given that pattern vectors from both the classes are Gaussianly distributed with unknown mean and covariances. In such a situation, first of all, it is necessary to estimate the unknown parameters $\emptyset$ with a supervised learning algorithm or an unsupervised learning algorithm (i.e., learning without a teacher).
In the situations when supervised learning is used, the Baye's rule is used for recursive estimation of the a posteriori probability density function $p(\Theta / X_1, \ldots, X_n)$ [Tou and Gonzalez (1974), Ho & Agarwala (1968)]. The computational feasibility of these recursive estimation algorithms depends critically on the Sufficient Statistic [Anderson (1958)] for the relevant prior and posterior density function. Raiffa and Schlaifer (1961) and Spragins (1965) have extensively studied the properties of prior and posterior density functions for which the recursive algorithms converge to yield the desired $\Theta$, in the limit. Prior and posterior density functions that satisfy this requirement are called conjugate or reproducing pairs.

In the special case when $p(X, C^i, \Theta)$ is Gaussian, i.e., $p(X|C^i, \Theta)$ is $N(\Theta^i, \Sigma)$ with given covariance matrix $\Sigma$ and $p(\Theta^i)$ is also normally distributed - it has been shown that $\Theta^i$ may be recursively estimated with a linear recursive equation [Abramson and Braverman (1962)]. In fact the problem of estimation of $\Theta^i$ under these conditions is equivalent to the Kalman - Bucy filtering problem in control theory [Kalman (1960)].

Some of the results that hold in the case of supervised learning have been extended for the cases of learning without a teacher.
[Meisel (1972)]. However, in the case of non-supervised learning, the convergence of the estimation algorithms is not known (in general) and very little computational implementation has been reported. In any of the above mentioned conditions, once $\Theta^1$ is estimated, the results of statistical decision theory may be applied for selecting $\gamma$. However, it may be noted that - given the optimal decision function as function of $\Theta$ and the best estimate of $\Theta$, it does not necessarily imply that the optimal decision function will still be optimal if $\Theta$ is replaced by its estimate.

Now we consider the most difficult situation where even the functional forms of the conditional density $p(X \mid C^1)$ is not known. Then it becomes necessary to estimate the density functions or estimate the discriminant function

$$f(X) = p(C^2 \mid X) = p(C^1 \mid X)$$

... (2.7)

Various methods for estimation of $f(X)$ (as in (2.7)) or $p(X \mid C^1)$, have been reported in the literature [Tsypkin (1966), Blaydon and Ho (1966), Kashyap and Blaydon (1966), Nicolic and Fu (1966), Greblicki (1978)]. In this situation $f(X)$ (or $p(X \mid C^1)$) is expressed as

$$f(X) = \sum_{j=0}^{\infty} \alpha_j \phi_j(X),$$

... (2.8)
where \( \phi_j(X) \) are some class of complete (possibly orthonormal) functions. The problem of estimation of \( f(X) \) then reduces to estimation of the unknown parameters \( \alpha_j \), using a set of samples of known classification. The parameters \( \alpha_j \) are estimated recursively using Stochastic approximation techniques. These methods of estimation are often called "potential function methods." [Arkadev and Braverman (1967)]. In practice, the convergence of most Stochastic approximation methods is extremely slow. Some time they appear not to converge at all. As suggested by Ho and Agrawala (1968) these techniques based on Stochastic approximation appear to be of theoretical interest only.

From the above discussion, some of the essential limitations of the statistical techniques may be outlined as follows. We have observed that the statistical methodologies might be effective if at least the form of the conditional density function is available for the classes under consideration. In such a case also if the form of the distribution is non-Gaussian, the estimation of any unknown parameters in the distribution function poses a difficult problem. Even if the form of the distribution function is known, the estimation of the unknown parameters requires that the training sample set is a sufficient statistic for the class of distribution.
function under consideration. It is however, difficult to ensure that any available training sample set is indeed a sufficient statistic for the selected class of distribution function. Above all, in practice, it is difficult to make any apriori guess about the most suitable form of the distribution function. The problem of estimation of the distribution function is tractable only when the form of the distribution function is known to be Gaussian. As shown in the last paragraph, if, even the form of the distribution function is not known, the techniques for estimation of the decision function do not yield a practically feasible solution. Hence, when sufficient statistical information is not available, the problem of estimation of suitable discriminant function becomes really a complex problem. It would be observed in the sequel that the fuzzy set theory, suggested by Zadeh (1965) provides a suitable framework for dealing with such problems.

2.2 Fuzzy Sets and Pattern Classification

In the classical set theory, the sets considered are defined as collection of objects having some very general property, nothing special is assumed or considered about the nature of the individual objects. For example, we can define a set A as the set of
all men. Symbolically

\[ A = \{ x | x \text{ is a man} \} \]

When we consider the "class of tall men", we can not define a classical set as the property "tall men" is not a precise property that may characterize the set. Similarly "the class of real numbers which are much greater than one" or "the class of beautiful women" do not constitute classes or sets in the usual mathematical sense of the term. In fact, most of the classes of object encountered in real life are of this "fuzzy" or "ill-defined type". They do not have a precisely defined criteria of membership. In such classes an object need not either necessarily belong or not belong to a class; there may be intermediate grades of membership. In order to deal with such classes, Zadeh (1965) has defined the "fuzzy sets" as classes with continuum of grades of membership. Essentially, such a framework provides a natural way of dealing with problems where the imprecision is due to vagueness rather than randomness.

**Fuzzy sets:**

We next formally define a fuzzy set and state some of its algebraic properties.

A Fuzzy subset \( A \) of a universe of discourse \( \mathbb{U} \), as proposed by Zadeh [1965], is characterized by a membership function
\[ \mu_A: \cap [0, 1] \text{which associates with each element in } x \text{ of } \cap \text{ a member } \mu_A(x) \text{ in the interval } [0, 1]. \] Thus a fuzzy subset \( A \) in \( \cap \) is a set of ordered pairs,

\[ A = \left\{ (x, \mu_A(x)) \mid x \in \Omega \right\}, \quad \cdots (2.9) \]

where \( \mu_A(x) \) represent the degree of membership of \( x \) in the fuzzy subset \( A \). Intuitively, a fuzzy subset \( A \) has an ill-defined boundary so that an element \( x \) is not necessarily either in \( A \) (with \( \mu_A(x) = 1 \)) or "not in \( A \" (with \( \mu_A(x) = 0 \)).

Let us consider \( \cap \) to be the real line and we consider the class of real numbers which are much greater than 1. We can define this set \( A \equiv \left\{ x \mid x \text{ is real number and } x \geq 1 \right\} \). As mentioned before, it is not a well defined set. It may be considered to be a fuzzy set

\[ A = \left\{ (x, \mu_A(x)) \right\}, \text{ where the membership function may be} \]

\[ \mu_A(x) = 0 \quad \text{for} \quad x \leq 1 \]

\[ \mu_A(x) = \frac{x - 1}{x} \quad \text{for} \quad x > 1. \]

\[ \cdots (2.10) \]

The assignment of the membership of a fuzzy set is subjective in nature and in general, reflects the context in which the problem is viewed [Kandel and Byatt (1978)]. However, it may be observed that though the assignment of the membership function of a fuzzy set is subjective,
it can not be assigned arbitrarily. For example, for the case mentioned above, it would be inadmissible to define $\mu_A$ to be:

$$
\mu_A(x) = \begin{cases} 
0 & \text{for } x \leq 1 \\
1 - e^{-x-1} & \text{for } x > 1 
\end{cases}
$$

... (2.11)

where $\mu_A(x)$ monotonically decreases as $x$ increases, for $x > 1$.

For the fuzzy set $A$, the membership function should be so chosen that it is consistent with the specification of the set. The example of another admissible membership function for the fuzzy set $A$ could be:

$$
\mu_A(x) = \begin{cases} 
0 & \text{for } x \leq 1 \\
1 - e^{-0.1(x-1)} & \text{for } x > 1 
\end{cases}
$$

... (2.12)

However, as indicated by Kandel and Byatt (1978), though the assignment of the membership function is subjective in nature yet it is quite logical to use any statistical information that is available. As an example, let us consider the fuzzy set of "tall men", where the height of a man is a statistical variable. While assigning the membership function for such a fuzzy set, the available statistics regarding the height of men may be used, such that if a man is less than 5 ft. tall (i.e., definitely short in common standard), then the membership function should assign him a low degree of membership in the fuzzy set of tall men.

As mentioned earlier, the fuzzy sets are essentially
generalization of the classical sets. Now we examine how some of the important set theoretic operations are interpreted in the realm of fuzzy sets. We present only the definitions of some basic operations [for a detailed discussion the reader may refer Zadeh (1965), Bellman and Gertz (1973), Zadeh et al (1975) and Kandel & Byatt (1978)].

**Definition 2.1** The union of two fuzzy sets A and B with respective membership functions \( \mu_A(x) \) and \( \mu_B(x) \) is a fuzzy set \( C \), whose membership function is related to those of \( A \) and \( B \) by

\[
\mu_C(x) = \max \left[ \mu_A(x), \mu_B(x) \right], \forall x \in \Omega \quad \ldots (2.13)
\]

**Definition 2.2** The intersection of two fuzzy sets \( A \) and \( B \) with respective membership functions \( \mu_A(x) \) and \( \mu_B(x) \) is a fuzzy set \( C \), whose membership function is related to those of \( A \) and \( B \) by

\[
\mu_C(x) = \min \left[ \mu_A(x), \mu_B(x) \right], \quad x \in \Omega \quad \ldots (2.14)
\]

**Definition 2.3** The complement of a fuzzy set \( A \), associated with the membership function \( \mu_A(x) \) is a fuzzy set \( A' \), where

\[
\mu_{A'}(x) = 1 - \mu_A(x), \quad x \in \Omega \quad \ldots (2.15)
\]

With union, intersection and complementation of fuzzy sets defined as above, it can be shown that the usual laws of associativity, distributivity
and De-Morgan's laws of classical set theory are also valid for fuzzy sets. However, if \( A' \) is the complement of fuzzy set \( A \), then unlike classical set theory \( A \cap A' \neq \emptyset \) and \( A \cup A' \neq \mathcal{U} \) where \( \emptyset \) and \( \mathcal{U} \) are respectively the null set and the universal set. In addition to the basic operations just defined, Zadeh (1973) has defined some additional operations which effectively create fuzzification on a set \( A \). The operations may be summarized as:

(i) Concentration of \( A \): \( \text{CON}(A) \). The operation implies

\[
\mu_{\text{CON}(A)}(x) = \left[ \mu_A(x) \right]^2, \quad \forall x \quad \ldots (2.16)
\]

(ii) Dilation of \( A \): \( \text{DIL}(A) \). The operation implies

\[
\mu_{\text{DIL}(A)}(x) = \left[ \mu_A(x) \right]^{0.5}, \quad \forall x \quad \ldots (2.17)
\]

(iii) Contrast intensification of \( A \): \( \text{INT}(A) \). The operation implies

\[
\mu_{\text{INT}(A)}(x) = \begin{cases} 
2 \left[ \mu_A(x) \right]^2, & 0 \leq \mu_A(x) \leq 0.5 \\
[1 - 2(1 - \mu_A(x))^2], & 0.5 \leq \mu_A(x) \leq 1.0
\end{cases} \quad \ldots (2.18)
\]

It may be noted that the notion of "belonging" which plays a fundamental role in the case of ordinary sets, does not have the same role in fuzzy sets. Thus it is not meaningful to speak of a point \( x \) "belonging" to a fuzzy set \( A \) except in the trivial sense of \( \mu_A(x) \) being positive. Less trivially, one can introduce two levels \( \alpha \) and \( \beta \)

\((0 < \alpha < 1, \quad 0 < \beta < 1, \quad \alpha > \beta)\) and agree to say that (a) "\( x \) belongs to \( A \)"

25
if $\mu_A(x) \geq \alpha$ (b) "x does not belong to A" if $\mu_A(x) \leq \beta$ and 
(c) "x has an indeterminate status relative to A" if $\beta < \mu_A(x) < \alpha$.

Pattern Classification

Patterns occurring in real life may be broadly classified into three categories: deterministic, probabilistic or fuzzy in nature. When we are faced with the design of a pattern classifier, usually we encounter patterns which are intrinsically probabilistic or fuzzy rather than deterministic. In the previous section we have discussed various techniques that might be used for dealing with patterns that are considered to be probabilistic and have also indicated that these techniques cannot be effectively applied when adequate statistical information is not available. Hence in the cases where the knowledge regarding the pattern classes is imprecise or where the patterns do not have a well-defined criteria of membership (because of vagueness rather than randomness), the use of fuzzy membership function and subsequent decision making appears to be quite attractive. In fact, fuzzy set theoretic approach to pattern classification has already found several applications


As mentioned by Bellman, Kalaba and Zadeh (1966), the pattern classification in the fuzzy environment critically depends on the
choice of a "good" membership function. In order to emphasize the importance of the choice of a good membership function, let us consider two fuzzy pattern classes $C_1$ and $C_2$ and a set of points $(X_1, \ldots, X_i, \ldots, X_n)$ and $(Y_1, \ldots, Y_j, \ldots, Y_m)$ known to be of the classes $C_1$ and $C_2$ respectively. Naturally any "good" membership function must satisfy:

$$
\mu_{C_1}(X_i) > \mu_{C_2}(X_i) \\
\text{and} \quad \mu_{C_2}(Y_j) > \mu_{C_1}(Y_j) 
$$

Essentially the problem of abstraction is estimation of a "good" membership function.

From (2.19), it may be observed that the membership function is an extension of the concept of discriminant function or likelihood function in the fuzzy environment. We have observed that linear discriminant function is often used in deterministic pattern classification techniques and linear discriminant function also arises out of minimum distance criterion for classification. If an appropriate metric be defined in the feature space, then the minimum distance criterion can also be used in designing the membership function. In fact, Pal and Dutta Majumdar (1977) and Dunn (1974) have defined membership functions using the Euclidean distance defined in the
feature space for solving the speech recognition problem and for identification of clusters in large data sets respectively. However, as suggested by Bellman, Kalaba and Zadeh (1966), the concept of membership function is more general and can be used even in the cases where an appropriate Euclidean metric cannot be defined in the feature space.

In order to estimate a "good" membership function, it is often necessary to have some apriori information about the class of functions to which \( \mu \) belongs, such that this information in combination with the samples from the pattern classes would be sufficient to obtain an estimate of the appropriate membership functions. As pointed out by Kandel and Byatt (1978), it would quite logical to use any statistical information regarding the pattern classes in defining the membership function. As observed in the statistical and deterministic pattern classification techniques, if the nature of the membership function is known the problem of estimation becomes tractable. So it would be desirable to assume the nature of the membership function with some unknown parameters. These unknown parameters may be estimated using a supervised or non-supervised learning algorithm. Pal, Datta and Dutta-Majumdar (1978) have studied the use of supervised as well as non-supervised learning algorithms for estimation of these parameters. In the next chapter we present a supervised learning algorithm for
Once the membership functions are estimated, a pattern vector $P$ is logically decided to be more likely member of one of $L$ fuzzy pattern classes $C_k$ if:

$$\mu_{C_k}(P) > \mu_{C_j}(P), \quad j = k = 1, 2, \ldots, L, \quad j \neq k$$

The linguistic classification technique based on fuzzy sets is briefly discussed in the next section.

2.3 **Linguistic (Syntactic) Methods in Pattern Classification**

The decision theoretic approach discussed in the previous sections is suited for applications where the patterns can be meaningfully represented in vector form. There are applications, however, where the structure of the pattern plays an important role in classification. As sentence structure is a principle object of study of natural and artificial languages, so pattern structure is the principle concern of linguistic pattern recognition. The methodologies of formal linguistics provide a convenient formalism for representing and operating on structural (syntactic) relationships. The two funda-
mental and indivisible aspects of the linguistic approach are describing the patterns of interest by formal grammars and the syntactic (linguistic) analysis of the linguistic representation of patterns.

Basic to the syntactic pattern recognition approach is the decomposition of patterns into subpatterns or primitives. Let us consider an example as shown in figure 2.1. The patterns under consideration are a square and right-triangle as shown in figure 2.1(b). The subpatterns (or primitives) are as shown in figure 2.1(a). The simplest form of representation is the string representation (as shown in figure 2). String representations are adequate for describing patterns whose structure is based on relatively simple connectivity of primitives. In general the strings are formed from the pattern primitives by simple operation of concatenation. The descriptive power of string representation may be increased by defining operators that allows higher - dimensional properties to be expressed in the form of a string. An example of use of such operators is found in the picture description language of Shaw (1969, 1970). Once the strings are formed using operators, the operators are also treated as pattern primitives. The string representation has been effectively used for solving various pattern recognition problems like Chromosome classification \[ \text{Ledley et al (1965)} \], Particle track detection in Bubble
Figure - 2.1(a)

Figure - 2.1(b)
When the patterns are represented as strings of primitives, they may be considered as sentences in a regular, context free, or context-sensitive languages. The next step is the construction of the grammar which will generate a language to describe the patterns under study. This is essentially the problem of grammatical inference from a finite set of sentences of the language. The problem of grammatical inference is the central problem in syntactical Pattern recognition system. Unfortunately there does not exist an algorithm for inference of any general class of grammars (for details see chapter 4 and 5 of the dissertation). By tracking the boundaries of these structures, in a clockwise direction, it is possible to describe the patterns as strings of primitives. Thus the square may be represented by the string (dddaaabbbccc) and the right triangle may be represented by the string (dddeeeecccc). The primitives may be considered to be the terminal symbols in some generative grammar. It is possible to envision two grammars $G_1$ and $G_2$ whose rules allow the generation of sentences that corresponds to the strings representing the right-triangle and square respectively. Thus the language $L(G_1)$ generated by $G_1$ would consist of sentences representing the right-triangles and
the language $L(G_2)$ generated by $G_2$ would consist of sentences representing squares. Once the two grammars $G_1$ and $G_2$ have been established, the linguistic pattern classification process is essentially the process of identifying the string of an input pattern (i.e., the sentence) to any of the languages under consideration. Thus for the above example, if the input sentence belongs to $L(G_1)$, the input pattern is classified as a right triangle; if it belongs to $L(G_2)$, it is classified as a square.

Effective linguistic analysis of patterns depends critically on the effective pattern representation. The problem of proper primitive selection and representation of the patterns in terms of the primitives may be viewed as the problem of characterization discussed earlier [Fu (1974), Gonzalez and Thomason (1978)]. The linguistic pattern classification techniques are essentially syntactic analysis of the patterns represented as strings of primitives. Though we are interested only in the techniques of pattern classification, it is impossible to ignore the linguistic description of patterns, as the linguistic analysis depends on the representation of the patterns. It may be observed that the increased descriptive power of a language is paid for in terms of increased complexity of the analysis system (recognizer). Finite state automata are capable of recognizing regular languages, although the descriptive power of finite state language is known to be weaker.
than that of context-free or context sensitive languages. On the other hand, non-finite, non-deterministic devices are required, in general, to accept the languages generated by context-free and context-sensitive grammars. Except for the class of deterministic languages (example, LR(K) languages and precedence languages) non-deterministic parsing procedures are usually required for context-free languages. The trade-off between the descriptive power and analysis efficiency of a grammar depends on the designer [Fu (1974), Fu (1977), Gonzales and Thomason (1978)]. Though the context-sensitive languages seem to be highly suitable for pattern description, yet due to the difficulty of analysis, the use of context-sensitive grammars has not found any application. Kanal (1972) has suggested the use of transformational grammars [Clowes (1969)] as an effective tool for description and analysis of pictorial patterns.

As mentioned earlier, once the grammars are constructed for each class of pattern, the classification phase consists of designing of recognizer for each class. In other words it is necessary to parse the input pattern with respect to each of the grammar under consideration. The classification of the input pattern is determined by the grammar with respect to which the string gets parsed. For regular grammars it is necessary to construct the corresponding
deterministic finite state automata. For the context-free grammars use of top-down/bottom-up parser, Early's parser or Cocke-Younger-Kasami Parsing algorithm have been reported in the literature [Aho and Ullman (1972), Lewis et al (1976), Gonzalez and Thomason (1978)]. So far we have discussed the simplest possible representation, the string representation. A more powerful approach for many applications is realized through the use of tree representation. Basically any hierarchical ordering scheme leads to a tree structure. The attractiveness of tree structures in linguistic pattern recognition is their ability to describe pattern primitives with multiple connectivity characteristics via nodes with more than one offspring. The grammars that are used to deal with tree structures are called tree grammars which are defined as generalization of string grammars [Fu (1974), 1977], Gonzalez and Thomason (1978)]. Some of the applications of tree grammars for pattern classification are -- Fingerprint classification [Moayer and Fu (1976)], Analysis of bubble chamber photograph [Fu and Bhargava (1973)], Recognition of Korean characters [Agui et al (1979)] etc. The automata that is used in classification of tree representation are called tree automata [Gonzalez and Thomason (1978)]. However, it has been reported that recognition problems for trees may be studied in terms of push down automata.
A further generalization of the pattern representation is by means of webs and plexes. Webs are indirected graphs. When they are used to represent syntactic structures, webs allow pattern description at a level considerably more general than that afforded by string or tree formalisms. The plexes are structures that include strings, trees and webs as substructures [Feder (1971)]. The web grammars were first introduced by Pfalz and Rosenfeld (1969). The notion of web automata has been introduced by Milgram (1972). Though these representation are quite general for representing higher dimensional structures, as far as practical pattern classification is concerned they have not received much application.

Till now we have considered classification of perfectly formed strings using recognizers that accept those strings. Since pattern processing systems often receive noisy or ill formed inputs for which a classification must be attempted, we now consider a technique for dealing with such non ideal situations.

The three types of error which are considered to corrupt the ideal pattern string are - the substitution error, where one terminal (usually representing a pattern primitive or relationship among primitives) is changed into another; the deletion error,
where one terminal is deleted; and an insertion error, where an extra terminal is inserted. For classification of strings with any one or more of the above-mentioned errors, various error correcting parsing schemes have been proposed [Aho and Peterson (1972), Lyon (1974), Fung and Fu (1974), Thompson (1979), Fu and Lu (1977), Tsai and Fu (1979)].

In our discussion thus far, attention has been focussed on non stochastic approaches to syntactic pattern classification. The different approaches discussed in the previous paragraphs, implicitly assumes that all the patterns under consideration are equally likely to occur. Hence, the classification schemes discussed above places the same degree of importance on all inputs classified into a particular pattern class. However, in some practical applications, a certain amount of uncertainty exists in the process of classification, as some patterns may occur more frequently than others and certain variations or distortions may be more likely than other deviations. In these cases the classification is often improved by employing probabilistic measures in the classification process. A given string \( x \) (corrupted by random error) may be classified as an element of more than one language and the appropriate classification may be ambiguous. To deal with such situations probabilities are assigned to the productions.
in the generative grammars in such a way that reflects the apriori
likelihoods of the classes and the individual sentences in them. Such
grammars, having probabilities associated with each production are
termed as stochastic or probabilistic grammars [Paz (1966),
Turakainen (1968), Fu and Li (1989), Fu (1974)]. Once the probabili-
ties are assigned to the productions in the grammars representing
each class, after the syntax analysis phase an input string (i.e., after
parsing) gets a probability measure associated to it with respect to
each of the grammars under consideration. Then statistical decision
theory is employed to make the final assignment of the string to a
pattern class. Various methods of estimation of production probabilities
and the use of stochastic syntax analysis for regular and context-free
grammars for pattern classification have been reported in the literature
[Fu (1974), Gonzalez and Thomason (1978), Maryanski and Thomason
(1977), Fung and Fu (1975), Pavlidis and Ali (1979)].

When the imprecision in the pattern strings is due to
vagueness rather than randomness, fuzzy, grammars and automata
[Thomason (1973), Lee and Zadeh (1969), Thomason and Marinos
(1974)], have been proposed for syntactic analysis. In a fuzzy
grammar a fuzzy membership value is attached to each production
in the grammar. The fuzzy pattern grammar produces a language,
which is a fuzzy set of strings with each string's membership in the
language is measured on the interval \([0, 1]\). The recognition scheme for fuzzy languages have been discussed by various authors Honda and Nasu (1975), Thomason and Marinos (1974), Wee and Fu (1969). Some of the applications of fuzzy syntactic analysis are - recognition handwritten English scripts [Depalma and Yau (1975)], Chromosome classification [Lee (1975)], recognition of handwritten capitals [Kickert and Koppeler (1976)], phonetic and phonemic labeling of continuous speech [Demori and Laface (1980)], etc.

Now that we have discussed the various approaches to pattern classification, in the next chapter we propose a hierarchical classification technique based on the theory of fuzzy sets.