Chapter 2
Review of Methodologies

The review of the previous chapter shows that country risk analysis is often based on the development of models to discriminate between high-risk countries (e.g., rescheduling) and low-risk ones.

This chapter is focused on the methods that can be used to develop such models. Special emphasis is given to non-parametric techniques from the field of MCDA. The considered methods include the UTADIS method and the MHDIS method (Multi-group Hierarchical DIScrimination). Both methods lead to the development of additive models that can be used to classify a set of alternatives (e.g., countries) into q predefined ordinal groups:

$$C_1 \succ C_2 \succ \cdots \succ C_q$$

where C_1 denotes the group consisting of the most preferred alternatives and C_q denotes the group of the least preferred alternatives. Within the country risk context, C_1 consists of the low-risk countries, whereas C_q consists of the high-risk ones.

The subsequent sections of this chapter discuss in detail all the model development aspects of the two methods as well as all the important issues of the model development and implementation process.

In addition, the two MCDA methods other techniques also discussed, including statistical methods, neural networks, rule induction and decision trees, fuzzy sets, and rough sets.

2.1 The UTADIS Method

2.1.1 Criteria Aggregation Model

The UTADIS method was first presented by Devaud et al. (1980), and some aspects of the method can also be found in Jacquet-Lagrèze and Siskos (1982). Jacquet-Lagrèze (1995) used the method to evaluate R & D projects, and during the past

K. Kosmidou et al. (eds.) *Country Risk Evaluation,* DOI: 10.1007/978-0-387-76680-5_2,
© Springer Science+Business Media, LLC 2008

few years the method has been widely used for developing classification models in financial decision making problems (Zopounidis and Doumpos, 1998, 1999a, b; Doumpos and Zopounidis, 1998; Zopounidis et al., 1999). Recently, the method has been implemented in multicriteria decision support systems, such as the FINCLAS system (Zopounidis and Doumpos, 1998) and the PREFDIS system (Zopounidis and Doumpos, 2000a).

The UTADIS method is a variant of the well-known UTA method (UTilités Additives). The latter is an ordinal regression method proposed by Jacquet-Lagrèze and Siskos (1982) for developing decision models that can be used to rank a set of alternatives from the best to the worst ones.

Within the sorting framework described in the introductory section of this chapter, the objective of the UTADIS method is to develop a criteria aggregation model used to determine the classification of the alternatives. Essentially this aggregation model constitutes an index representing the overall performance of each alternative along all criteria. The objective of the model development process is to specify this model so that the alternatives of group C_1 receive the highest scores, while the scores of the alternatives belonging to other groups gradually decrease as we move toward the worst group C_q.

Formally, the criteria aggregation model is expressed as an additive utility function:

$$U(\mathbf{g}) = \sum_{i=1}^{n} p_i u_i(g_i) \tag{2.1}$$

where:

$\mathbf{g} = (g_1, g_2, \ldots, g_n)$ is the vector of the evaluation criteria.

p_i is a positive scaling constant indicating the significance of criterion g_i ($p_1 + p_2 + \cdots + p_n = 1$).

$u_i(g_i)$ is the marginal utility function of criterion g_i.

The marginal utility functions are monotone functions (linear or nonlinear) defined on the criteria's scale, such that the following two conditions are met:

$$\left.\begin{array}{l} u_i(g_{i*}) = 0 \\ u_i(g_i^*) = 1 \end{array}\right\}$$

where g_{i*} and g_i^* denote the least and the most preferred value of criterion g_i, respectively. These values are specified according to the set of the alternatives under consideration, as follows:

- For increasing preference criteria (criteria for which higher values indicate higher preference, e.g., return/profitability criteria):

$$g_{i*} = \min_{\forall \mathbf{x}_j \in A} \{g_{ji}\} \quad \text{and} \quad g_i^* = \max_{\forall \mathbf{x}_j \in A} \{g_{ji}\}$$

- For decreasing preference criteria (criteria for which higher values indicate lower preference, e.g., risk/cost criteria):

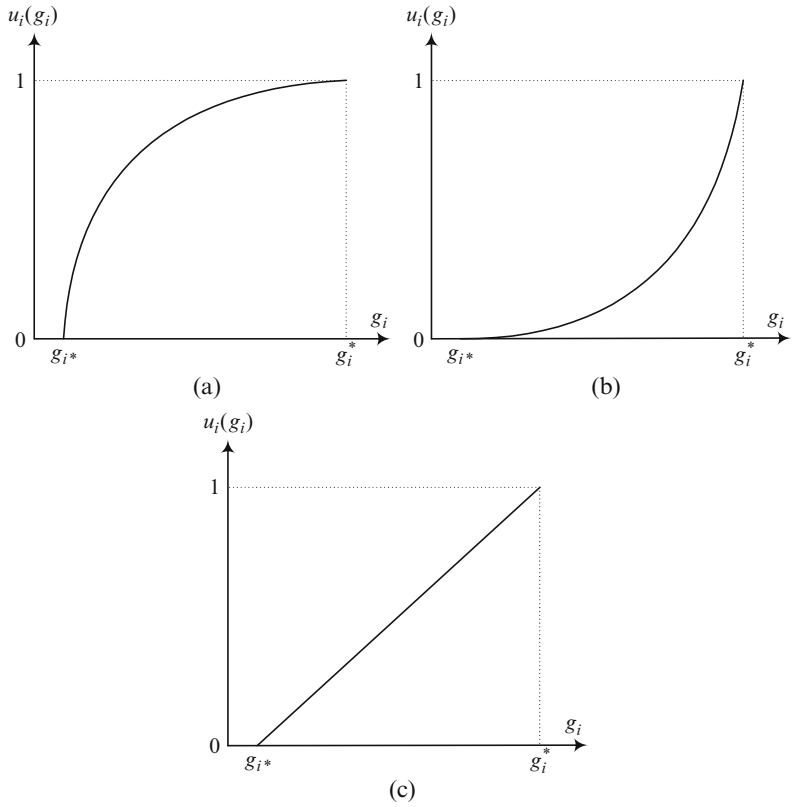

Fig. 2.1 Characteristic forms of marginal utility functions

$$g_{i*} = \max_{\forall \mathbf{x}_j \in A} \{g_{ji}\} \quad \text{and} \quad g_i^* = \min_{\forall \mathbf{x}_j \in A} \{g_{ji}\}$$

Essentially, the marginal utility functions provide a mechanism for transforming the criterion's scale into a new scale ranging in the interval $[0, 1]$. This new scale represents the utility for the decision maker of each value of the criterion. The form of the marginal utility functions depends upon the decision maker's preferential system (judgment policy). Figure 2.1 presents three characteristic cases. The concave form of the utility function presented in Figure 2.1(a) indicates that the decision maker considers as quite significant small deviations from the worst performance g_i*. This corresponds with a risk-averse attitude. On the contrary, the case presented in Figure 2.1(b) corresponds with a risk-prone decision maker who is mainly interested in alternatives of top performance. Finally, the linear marginal utility function of Figure 2.1(c) indicates a risk-neutral behavior.

Transforming the criteria's scale into utility terms through the use of marginal utility functions has two major advantages:

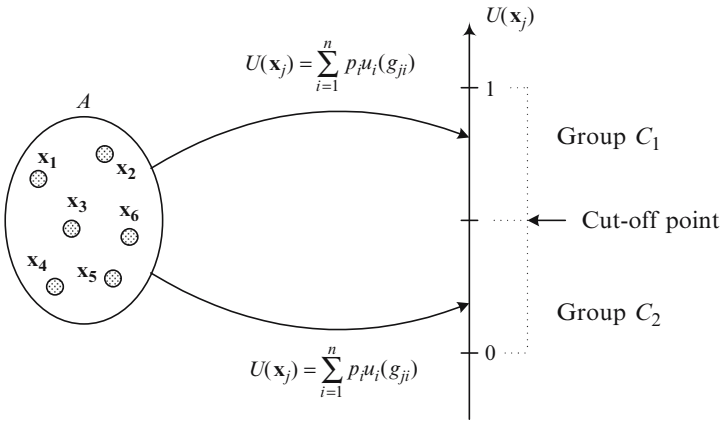

Fig. 2.2 Classification of the alternatives on the basis of their global utilities

1. It enables the modeling and representation of the nonlinear behavior of the decision maker when evaluating the performance of the alternatives.
2. It enables the consideration of qualitative criteria in a flexible way.

Given the above discussion on the concept of marginal utilities, the global utility of an alternative $\mathbf{x}_j$ specified through eq. (2.1) represents a measure of the overall performance of the alternative considering its performance on all criteria. The global utilities range in the interval [0, 1] and they constitute the criterion used to decide upon the classification of the alternatives. Figure 2.2 illustrates how the global utilities are used for classification purposes in the simple two group case. The classification is performed by comparing the global utility of each alternative with a cutoff point defined on the utility scale between 0 and 1. Alternatives with global utilities higher than the utility cutoff point are assigned into group C_1, whereas alternatives with global utilities lower than the cutoff point are assigned into group C_2.

In the general case where q groups are considered, the classification of the alternatives is performed through the following classification rules:

$$\left.\begin{array}{rl} U(\mathbf{x}_j) \geq u_1 & \Rightarrow \mathbf{x}_j \in C_1 \\ u_2 \leq U(\mathbf{x}_j) < u_1 & \Rightarrow \mathbf{x}_j \in C_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\ U(\mathbf{x}_j) < u_{q-1} & \Rightarrow \mathbf{x}_j \in C_q \end{array}\right\} \tag{2.2}$$

where $u_1, u_2, \ldots, u_{q-1}$ denote the utility cutoff points separating the group. Henceforth, these cutoff points will be referred to as utility thresholds. Essentially, each utility threshold u_k separates two consecutive groups C_k and C_{k+1}.

2.1.2 Model Development Process

2.1.2.1 General Framework

The main structural parameters of the classification model developed through the UTADIS method include the criteria weights, the marginal utility functions, and the utility thresholds. These parameters are specified through the regression-based philosophy of preference disaggregation analysis.

A general outline of the model development procedure in the UTADIS method is presented in Figure 2.3.

Initially, a reference set A' consisting of m alternatives described along n criteria is used as the training sample (henceforth the training sample will be referred to as the *reference set* in order to comply with the terminology used in MCDA). The alternatives of the reference set are classified a priori into q groups. The reference set should be constructed in such a way so that it includes an adequate number of representative examples (alternatives) from each group. Henceforth, the number of alternatives of the reference set belonging to group C_k will be denoted by m_k.

Given the classification C of the alternatives in the reference set, the objective of the UTADIS method is to develop a criteria aggregation model and a set of utility thresholds that minimize the classification error rate. The error rate refers to the differences between the estimated classification $\widehat{C}$ defined through the developed model and the prespecified classification C for the alternatives of the reference set. Such differences can be represented by introducing a binary variable E representing the classification status of each alternative:

$$E_j = \begin{cases} 0, & \text{if } \mathbf{x}_j \text{ is correctly classified} \\ 1, & \text{if } \mathbf{x}_j \text{ is misclassified} \end{cases}$$

On the basis of this binary variable, the classification error rate γ is defined as the ratio of the number of misclassified alternatives to the total number of alternatives in the reference set:

$$\gamma = \frac{\sum_{j=1}^{m} E_j}{m} \in [0, 100\%] \tag{2.3}$$

This classification error rate measure is adequate for cases where the number of alternatives of each group in the reference set is similar along all groups (i.e., $m_1 \approx m_2 \approx \cdots \approx m_q$). In the case, however, where there are significant differences, then the use of the classification error rate defined in (2.3) may lead to misleading results. For instance, consider a reference set consisting of 10 alternatives, 7 belonging into group C_1 and 3 belonging into group C_2 ($m_1 = 7$, $m_2 = 3$). In this case, a classification that assigns correctly all alternatives of group C_1 and incorrectly all alternatives of group C_2 has an error rate $\gamma = 30\%$. This is a misleading result. Actually, what should be the main point of interest in the expected classification error $\Pr(error)$. This is expressed in relation to the a priori probabilities π_1 and π_2 that an alternative belongs to groups C_1 and C_2, respectively, as follows:

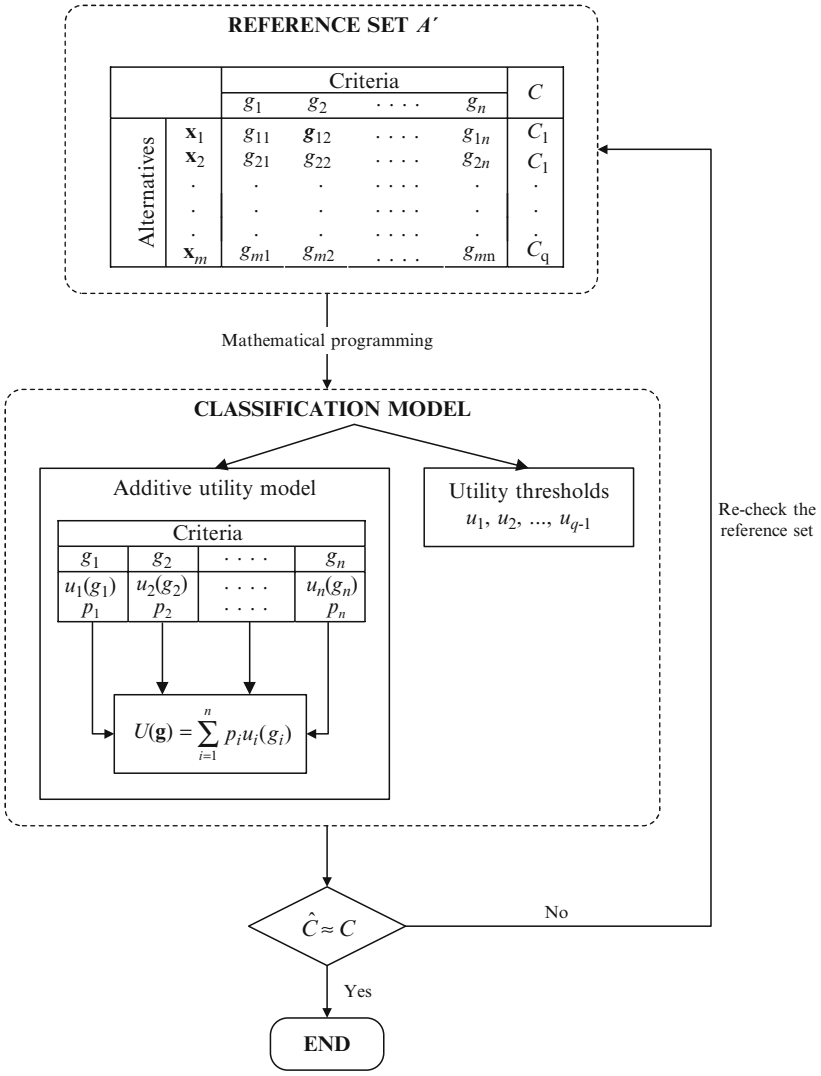

Fig. 2.3 Outline of the model development procedure in the UTADIS method

$$\begin{aligned}\Pr(error) &= \Pr(\text{incorrect classification of an alternative } \mathbf{x}_j)\\ &= \Pr[(\mathbf{x}_j \in C_1 \text{ and assigned in } C_2) \text{ or } (\mathbf{x}_j \in C_2 \text{ and assigned in } C_1)]\\ &= \Pr[(\mathbf{x}_j \in C_1) \wedge (\mathbf{x}_j \to C_2)] + \Pr[(\mathbf{x}_j \in C_2) \wedge (\mathbf{x}_j \to C_1)]\\ &= \pi_1 \Pr(\mathbf{x}_j \to C_2) + \pi_2 \Pr(\mathbf{x}_j \to C_1)\end{aligned}$$

In the above example, the error rates for the two groups (0% for C_1 and 100% for C_2) can be considered as estimates for the probabilities $\Pr(\mathbf{x}_j \to C_1)$ and $\Pr(\mathbf{x}_j \to C_2)$, respectively. Assuming that the a priori probabilities for the two groups are equal (i.e., $\pi_1 = \pi_2 = 0.5$), then the expected error of the classification is 0.5. This result indicates that the obtained classification corresponds with a random

classification. In a random classification, the probabilities $\Pr(\mathbf{x}_j \to C_k)$ are determined based on the proportion of each group C_k to the total number of alternatives in the reference set. In this respect, in the above example a naïve approach would be to assign 7 out of the 10 alternatives into group C_1, i.e., $\Pr(\mathbf{x}_j \to C_1) = 0.7$, and 3 out of the 10 alternatives into group C_2, i.e., $\Pr(\mathbf{x}_j \to C_2) = 0.3$. The expected error of such a naïve approach (random classification) is 0.5.

To overcome this problem, a more appropriate measure of the expected classification error rate is expressed as follows:

$$\gamma = \sum_{k=1}^{q} \left(\pi_k \frac{\sum\limits_{\forall \mathbf{x}_j \in C_k} E_j}{m_k} \right) \in [0, 100\%] \tag{2.4}$$

Even though this measure takes into consideration the a priori probabilities of each group, it assumes that all classification errors are of equal cost to the decision maker. This is not always the case. For instance, the classification error regarding the assignment of a bankrupt firm to the group of healthy firms is much more costly than an error involving the assignment of a healthy firm to the bankrupt group. The former leads to capital cost (loss of the amount of credit granted to a firm), whereas the latter leads to opportunity cost (loss of profit that would result from granting a credit to a healthy firm). Therefore, it would be appropriate to extend the expected classification error rate (2.4) so that the costs of each individual error are also considered. The resulting measure represents the expected misclassification cost (EMC), rather than the expected classification error rate:

$$\text{EMC} = \sum_{k=1}^{q} \left[\pi_k \left(\sum_{\substack{l=1 \\ l \neq k}}^{q} K_{kl} \sum_{\forall \mathbf{x}_j \in C_k} \frac{E_{klj}}{m_k} \right) \right] \in [0, 1] \tag{2.5}$$

where:

- K_{kl} is the misclassification cost involving the classification of an alternative of group C_k into group C_l ($l \neq k$).
- E_{klj} is a binary $0-1$ variable defined such that $E_{klj} = 1$ if an alternative $\mathbf{x}_j \in C_k$ is classified into group C_l ($l \neq k$) and $E_{klj} = 0$ if $\mathbf{x}_j$ is not classified into group C_l.

Comparing expressions (2.4) and (2.5), it becomes apparent that the expected classification error rate in (2.4) is a special case of the expected misclassification cost, when all costs K_{kl} are considered equal for every $k, l = 1, 2, \ldots, q$. The main difficulty related to the use of the expected misclassification cost as the appropriate measure of the quality of the obtained classification is that it is often quite difficult to have reliable estimates for the cost of each type of classification error.

If the expected classification error rate, regarding the classification of the alternatives that belong into the reference set, is considered satisfactory, then this is an indication that the developed classification model might be useful in providing reliable recommendations for the classification of other alternatives. On the other hand,

if the obtained expected classification error rate indicates that the classification of the alternatives in the reference set is close to a random classification (i.e., $\gamma \approx 1/q$ or $\gamma > 1/q$), then the decision maker must check the reference set regarding its completeness and adequacy for providing representative information on the problem under consideration. Alternatively, it is also possible that the criteria aggregation model (additive utility function) is not able to provide an adequate representation of the decision maker's preferential system. In such a case, an alternative criteria aggregation model must be considered.

However, it should be pointed out that a low expected classification error rate does not necessarily ensure the practical usefulness of the developed classification model; it simply provides an indication supporting the possible usefulness of the model. On the contrary, a high expected classification error rate leads with certainty to the conclusion that the developed classification model is inadequate.

2.1.2.2 Mathematical Formulation

Pursuing the objective of the model development process in the UTADIS method, i.e., the maximization of the consistency between the estimated classification $\widehat{C}$ and the predefined one C, is performed through mathematical programming techniques.

In particular, the minimization of the expected classification error rate (2.4) requires the formulation and solution of a mixed-integer programming (MIP) problem. The solution, however, of MIP formulations is a computationally intensive procedure. Despite the significant research that has been made on the development of computationally efficient techniques for solving MIP problems within the context of classification model development, the computational effort still remains quite significant. This problem is most significant in cases where the reference set includes a large number of alternatives.

To overcome this problem, an approximation of the error rate (2.4) is used as follows:

$$\gamma' = \frac{1}{q} \sum_{k=1}^{q} \left(\frac{\sum_{\forall \mathbf{x}_j \in C_k} \sigma_j}{m_k} \right) \tag{2.6}$$

where σ_j is a positive real variable, defined such that:

$$\sigma_j = \begin{cases} > 0, & \text{if } \mathbf{x}_j \text{ is misclassified} \\ 0, & \text{if } \mathbf{x}_j \text{ is classified correctly} \end{cases} \tag{2.7}$$

Essentially, σ_j represents the magnitude of the classification error for alternative $\mathbf{x}_j$. On the basis of the classification rule (2.2), the classification error for an alternative of group C_1 involves the violation of the utility threshold u_1 that defines the lower bound of group C_1. For the alternatives of the last (least preferred) group C_q, the classification error involves the violation of the utility threshold u_{q-1} that defines the upper bound of group C_q. For any other intermediate group C_k $(1 < k < q)$, the

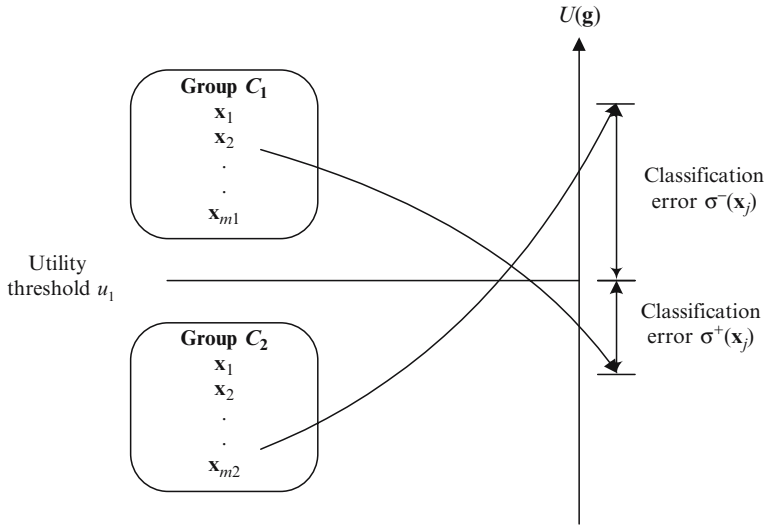

Fig. 2.4 The classification errors (two-group case)

classification error may involve either the violation of the upper bound of the group (utility threshold u_{k-1}) or the violation of the lower bound u_k.

Henceforth, the violation of the lower bound of a group will be denoted by σ^+, whereas σ^- will be used to denote the violation of the upper bound of a group. Figure 2.4 provides a graphical representation of these two errors in the simple two-group case. By definition, it is not possible that the two errors occur simultaneously (i.e., $\sigma^+\sigma^- = 0$). Therefore, the total error σ_j for an alternative $\mathbf{x}_j$ is defined as $\sigma_j = \sigma_j^+ + \sigma_j^-$.

At this point, it should be emphasized that the error functions (2.4) and (2.6) are not fully equivalent. For instance, consider a reference set consisting of four alternatives classified into two groups: $\{\mathbf{x}_1, \mathbf{x}_2\} \in C_1$, $\{\mathbf{x}_3, \mathbf{x}_4\} \in C_2$. Assume that for this reference set an additive utility classification model (CM1) is developed that misclassifies alternatives $\mathbf{x}_2$ and $\mathbf{x}_4$, such that $\sigma_2^+ = 0.2$ and $\sigma_4^- = 0.1$. Then according to (2.6) the total classification error is $\gamma' = 0.075$, whereas considering (2.4) the expected classification error rate is $\gamma = 50\%$. An alternative classification model (CM2) that classifies correctly $\mathbf{x}_2$ but retains the misclassification of $\mathbf{x}_4$ such that $\sigma_4^- = 0.5$ has $\gamma' = 0.125$ and $\gamma = 25\%$. Obviously, the model CM1 outperforms CM2 when the definition (2.6) is considered, but according to the expected classification error rate (2.4) CM2 performs better.

Despite this limitation, the definition (2.6) provides a good approximation of the expected classification error rate (2.4) while reducing the computational effort required to obtain an optimal solution.

The two forms of the classification errors can be formally expressed on the basis of the classification rule (2.2) as follows:

$$\sigma_j^+ = \max\{0, u_k - U(\mathbf{g}_j)\}, \qquad \forall \mathbf{x}_j \in C_k, k = 1,2,\ldots,q-1$$
$$\sigma_j^- = \max\{0, U(\mathbf{g}_j) - u_{k-1}\}, \qquad \forall \mathbf{x}_j \in C_k, k = 2,3,\ldots,q$$

These expressions illustrate better the notion of the errors. The error σ_j^+ indicates that to classify correctly a misclassified alternative $\mathbf{x}_j$ that actually belongs in group C_k, its global utility $U(\mathbf{x}_j)$ should be increased by $u_k - U(\mathbf{x}_j)$. Similarly, the σ_j^- indicates that to classify correctly a misclassified alternative $\mathbf{x}_j$ that actually belongs in C_k, its global utility $U(\mathbf{x}_j)$ should be decreased by $U(\mathbf{x}_j) - u_{k-1}$.

Introducing the error terms in the additive utility model, it is possible to rewrite the classification rule (2.2) in the form of the following constraints:

$$U(\mathbf{g}_j) + \sigma_j^+ \geq u_1, \qquad \forall \mathbf{x}_j \in C_1 \tag{2.8}$$
$$U(\mathbf{g}_j) + \sigma_j^+ \geq u_k, \qquad \forall \mathbf{x}_j \in C_k \ (k = 2,\ldots,q-1) \tag{2.9}$$
$$U(\mathbf{g}_j) - \sigma_j^- < u_{k-1}, \qquad \forall \mathbf{x}_j \in C_k \ (k = 2,\ldots,q-1) \tag{2.10}$$
$$U(\mathbf{g}_j) - \sigma_j^- < u_{q-1}, \qquad \forall \mathbf{x}_j \in C_{q-1} \tag{2.11}$$

These constraints constitute the basis for the formulation of a mathematical programming problem used to estimate the parameters of the additive utility classification model (utility thresholds, marginal utilities, criteria weights). The general form of this mathematical programming model is the following (MP):

$$\min \quad \sum_{k=1}^{q} \left[\frac{\sum_{\forall \mathbf{x}_j \in C_k} \left(\sigma_j^+ + \sigma_j^- \right)}{m_k} \right] \tag{2.12}$$
$$\text{s.t.} \quad U(\mathbf{g}_j) - u_1 + \sigma_j^+ \geq \delta_1, \qquad \forall \mathbf{x}_j \in C_1 \tag{2.13}$$
$$U(\mathbf{g}_j) - u_k + \sigma_j^+ \geq \delta_1, \qquad \forall \mathbf{x}_j \in C_k \ (k = 2,3,\ldots,q-1) \tag{2.14}$$
$$U(\mathbf{g}_j) - u_{k-1} - \sigma_j^- \leq -\delta_2, \qquad \forall \mathbf{x}_j \in C_k \ (k = 2,3,\ldots,q-1) \tag{2.15}$$
$$U(\mathbf{g}_j) - u_{q-1} - \sigma_j^- \leq -\delta_2, \qquad \forall \mathbf{x}_j \in C_q \tag{2.16}$$
$$U(\mathbf{g}^*) = 1 \tag{2.17}$$
$$U(\mathbf{g}_*) = 0 \tag{2.18}$$
$$u_k - u_{k+1} \geq s, \qquad k = 1,2,\ldots,q-1 \tag{2.19}$$
$$u_i(g_i) \text{ increasing functions} \tag{2.20}$$
$$\sigma_j^+, \sigma_j^- \geq 0, \qquad j = 1,2,\ldots,m \tag{2.21}$$

In constraints (2.13)–(2.14), δ_1 is a positive constant used to avoid cases where $U(\mathbf{g}_j) = u_k$ when $\mathbf{x}_j \in C_k$. Of course, u_k is considered as the lower bound of group C_k. In this regard, the case $\delta_1 = 0$, typically, does not pose any problem during model development and implementation. However, assuming the simple two-group case, the specification $\delta_1 = 0$ may lead to the development of a classification model for which $U(\mathbf{g}_j) = u_1 = 1$, for all $\mathbf{x}_j \in C_1$, and $U(\mathbf{g}_j) < u_1 = 1$ for all $\mathbf{x}_j \in C_2$. Because

the utility threshold u_1 is defined as the lower bound of group C_1, it is obvious that such a model performs an accurate classification of the alternatives. Practically, however, because all alternatives of group C_1 are placed on the utility threshold, the generalizing ability of such a model is expected to be limited. Therefore, to avoid such situations, a small positive (non-zero) value for the constant δ_1 should be chosen. The constant δ_2 in (2.15)–(2.16) is used in a similar way.

Constraints (2.17) and (2.18) are used to normalize the global utilities in the interval $[0,1]$. In these constraints, $\mathbf{g}_*$ and $\mathbf{g}^*$ denote the vectors consisting of the least and the most preferred levels of the evaluation criteria. Finally, constraint (2.19) is used to ensure that the utility threshold u_k is higher than the utility threshold u_{k+1}, thus ensuring the ordering of the groups from the most preferred (C_1) to the least preferred ones (C_q). In this ordering of the groups, higher utilities are assigned to the most preferred groups. In constraint (2.19), s is a constant defined such that $s > \delta_1, \delta_2$.

Introducing the additive utility function (2.1) in MP leads to the formulation of a nonlinear programming problem. This is because the additive utility function (2.1) has two unknown parameters: (a) the criteria weights and (b) the marginal utility functions. Therefore, constraints (2.13)–(2.18) take a nonlinear form, and the solution of the resulting nonlinear programming problem can be cumbersome. To overcome this problem, the additive utility function (2.1) is rewritten in a simplified form as follows:

$$U(\mathbf{g}) = \sum_{i=1}^{n} u_i'(g_i) \tag{2.22}$$

where:

$$\left.\begin{array}{l} u_i'(g_i) = p_i u_i(g_i) \\ u_i'(g_{i*}) = 0 \\ u_i'(g_i^*) = p_i \end{array}\right\} \tag{2.23}$$

Both (2.1) and (2.22) are equivalent expressions for the additive utility function. Nevertheless, the latter requires only the specification of the marginal utility functions $u_i'(g_i) \in [0, p_i]$. As illustrated in Figure 2.1, these functions can be of any form. The UTADIS method does not prespecify a functional form for these functions. Therefore, it is necessary to express the marginal utility functions in terms of specific decision variables to be estimated through the solution of MP. This is achieved through the modeling of the marginal utilities as piece-wise linear functions through a process that is graphically illustrated in Figure 2.5.

The range $[g_{i*}, g_i^*]$ of each criterion is divided into $a_i - 1$ subintervals $[g_i^h, g_i^{h+1}]$, $h = 1, 2, \ldots, a_i - 1$. The estimation of the unknown marginal utility functions can be performed by estimating the marginal utilities at the break-points $g_i^2, \ldots, g_i^{a_i}$. As illustrated in Figure 2.5, this estimation provides an approximation of the true marginal utility functions. On the basis of this approach, it would be reasonable to assume that the larger the number of subintervals that are specified, the better is the approximation of the marginal utility functions. The definition of a large number of subintervals, however, provides increased degrees of freedom to the additive utility model. This increases the fitting ability of the developed model to the data of

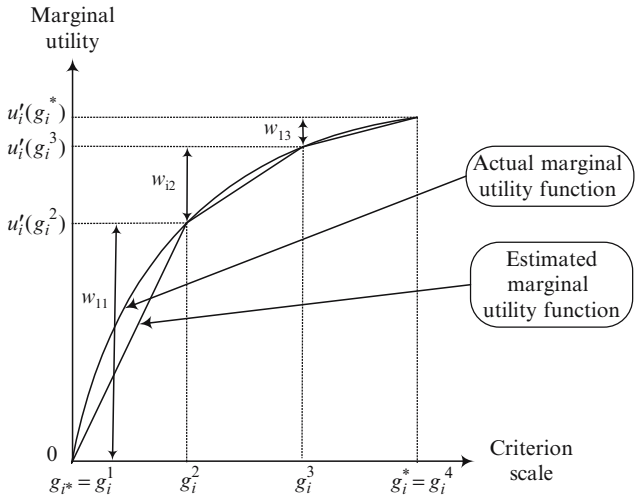

Fig. 2.5 Piece-wise linear form of marginal utility functions

the reference set; the instability, however, of the model is also increased (the model becomes sample-based).

The marginal utility at the break-point g_i^h is written as follows:

$$u_i'(g_i^h) = \sum_{t=1}^{h-1} w_{it}$$

where $w_{it} = u'(g_i^t) - u'(g_i^{t-1}) \geq 0$ are the parameters that must be estimated in order to specify the marginal value function. With this modeling, the marginal value function of any alternative $\mathbf{x}_j$ on the criterion g_i is expressed as follows:

$$u'(g_{ji}) = \sum_{t=1}^{r_{ji}-1} w_{it} + \frac{g_{ji} - g_i^{r_{ji}}}{g_i^{r_{ji}+1} - g_i^{r_{ji}}} w_{i,r_{ji}}$$

where r_{ji} $(1 \leq r_{ji} \leq a_i - 1)$ denotes the subinterval $[g_i^{r_{ji}},\ g_i^{r_{ji}+1}]$ into which the performance g_{ji} of alternative $\mathbf{x}_j$ on criterion g_i belongs to. The global utility of the alternative $\mathbf{x}_j$ is also expressed in terms of the unknown parameters w:

$$U(\mathbf{g}_j) = \sum_{i=1}^{n} \left(\sum_{t=1}^{r_{ji}-1} w_{it} + \frac{g_{ji} - g_i^{r_{ji}}}{g_i^{r_{ji}+1} - g_i^{r_{ji}}} w_{i,r_{ji}} \right)$$

Therefore, the problem MP is explicitly written as the following linear programming problem (LP):

$$\min \sum_{k=1}^{q} \left[\frac{\sum_{\forall \mathbf{x}_j \in C_k} \left(\sigma_j^+ + \sigma_j^- \right)}{m_k} \right] \tag{2.24}$$

$$\text{s.t.} \sum_{i=1}^{n} \left(\sum_{t=1}^{r_{ji}-1} w_{it} + \frac{g_{ji} - g_i^{r_{ji}}}{g_i^{r_{ji}+1} - g_i^{r_{ji}}} w_{i,r_{ji}} \right) - u_1 + \sigma_j^+ \geq \delta_1, \qquad \forall \mathbf{x}_j \in C_1 \tag{2.25}$$

$$\sum_{i=1}^{n} \left(\sum_{t=1}^{r_{ji}-1} w_{it} + \frac{g_{ji} - g_i^{r_{ji}}}{g_i^{r_{ji}+1} - g_i^{r_{ji}}} w_{i,r_{ji}} \right) - u_k + \sigma_j^+ \geq \delta_1, \qquad \forall \mathbf{x}_j \in \{C_2, \ldots, C_{q-1}\} \tag{2.26}$$

$$\sum_{i=1}^{n} \left(\sum_{t=1}^{r_{ji}-1} w_{it} + \frac{g_{ji} - g_i^{r_{ji}}}{g_i^{r_{ji}+1} - g_i^{r_{ji}}} w_{i,r_{ji}} \right) - u_{k-1} - \sigma_j^- \leq -\delta_2, \quad \forall \mathbf{x}_j \in \{C_2, \ldots, C_{q-1}\} \tag{2.27}$$

$$\sum_{i=1}^{n} \left(\sum_{t=1}^{r_{ji}-1} w_{it} + \frac{g_{ji} - g_i^{r_{ji}}}{g_i^{r_{ji}+1} - g_i^{r_{ji}}} w_{i,r_{ji}} \right) - u_{q-1} - \sigma_j^- \leq -\delta_2, \quad \forall \mathbf{x}_j \in C_q \tag{2.28}$$

$$\sum_{i=1}^{n} \sum_{t=1}^{a_i-1} w_{it} = 1 \tag{2.29}$$

$$u_k - u_{k-1} \geq s, \qquad 1 \leq k \leq q-1 \tag{2.30}$$

$$w_{it}, \sigma_j^+, \sigma_j^- \geq 0, \qquad \forall j, i, t \tag{2.31}$$

Constraints (2.25)–(2.28), and (2.29)–(2.30) correspond with the constraints (2.13)–(2.16), (2.17), and (2.19) of MP. The non-negativity constraint on the variables w ensures that the marginal value functions are increasing (constraint (2.20) in MP).

2.1.3 Model Development Issues

The simple linear form of LP ensures the existence of a global optimum solution. However, often there are multiple optimal solutions. The existence of multiple optimal solutions is most often when the groups are perfectly separable, i.e., when there is no group overlap. In such cases, all error variables σ_j^+ and σ_j^- are zero. The determination of a large number of criteria subintervals is positively related to the existence of multiple optimal solutions (as already mentioned, as the number of subintervals increases, the degrees of freedom of the developed additive utility model also increases and so does the fitting ability of the model).

In addition to the above phenomenon, it is also important to emphasize that even if a unique optimal solution does exist for LP, its stability needs to be carefully considered. A solution is considered to be stable if it is not significantly affected by small trade-offs to the objective function (i.e., if near-optimal solutions are

quite similar to the optimal one). The instability of the optimal solution is actually the result of overfitting the developed additive utility model to the alternatives of the reference set. This may affect negatively the generalizing classification performance of the developed classification model. In addition to the classification performance issue, the instability of the additive utility model also raises interpretation problems. If the developed model is unstable, then it is clearly very difficult to derive secure conclusions on the contribution of the criteria in the classification of the alternatives (the criteria weights are unstable and therefore difficult to interpret).

The consideration of these issues in the UTADIS method is performed through a post-optimality analysis that follows the solution of LP. The objective of post-optimality analysis is to explore the existence of alternate optimal solutions and near-optimal solutions. There are many different ways that can be used to perform the post-optimality stage considering the parameters that are involved in the model development process. These parameters include the constants δ_1, δ_2, and s, as well as the number of criteria subintervals. The use of mathematical programming techniques provides increased flexibility in considering a variety of different forms for the post-optimality analysis. Some issues that are worth the consideration in the post-optimality stage include:

1. The maximization of the constants δ_1 and δ_2. This implies a maximization of the minimum distance between the correctly classified alternatives and the utility thresholds, thus resulting in a more clear separation of the groups.
2. Maximization of the sum of the differences between the global utilities of the correctly classified alternatives from the utility thresholds. This approach extends the previous point considering all differences instead of the minimum ones.
3. Minimization of the total number of misclassified alternatives using the error function (2.4).
4. Determination of the minimum number of criteria subintervals.

Considering, however, the issues regarding the stability of the developed model and its interpretation, none of these approaches ensures the existence of a unique and stable solution. Consequently, the uncertainty on the interpretation of the model is still an issue to be considered.

To overcome this problem, the post-optimality stage performed in the UTADIS method focuses on the investigation of the stability of the criteria weights rather than on the consideration of the technical parameters of the model development process. In particular, during the post-optimality stage $n+q-1$ new linear programs are solved, each having the same form with LP. The solution of LP2 is used as input to each of these new linear programs to explore the existence of other optimal or near-optimal solutions. The objective function of each problem s involves the maximization of each criterion weight (for $s = 1,2,\ldots,n$) and the value of the utility thresholds (for $s > n$). All new solutions found during the post-optimality stage are optimal or near optimal for LP. This is ensured by imposing the following constraint:

$$f' \leq (1+z)f^*$$

where:

- f^* is the optimal value for the objective function of LP,
- f' is the value of the objective function of LP evaluated for any new solution obtained during the post-optimality stage.
- z is a small portion of f^* (a trade-off made to the optimal value of the objective function in order to investigate the existence of near-optimal solutions).

This constraint in added to the formulation of LP, and the new linear program that is formed is solved to maximize either the criteria weights or the utility thresholds as noted above. Finally, the additive utility model used to perform the classification of the alternatives is formed from the average of all solutions obtained during the post-optimality stage.

Overall, despite the problems raised by the existence of multiple optimal solutions, it should be noted that LP provides consistent estimates for the parameters of the additive utility classification model. The consistency property for mathematical programming formulations used to estimate the parameters of a decision-making model was first introduced by Charnes et al. (1955). The authors consider a mathematical programming formulation to satisfy the consistency property if it provides estimates of the model's parameters that approximate (asymptotically) the true values of the parameters as the number of observations (alternatives) used for model developed increases. According to the authors, this is the most significant property that a mathematical programming formulation used for model development should have, as it ensures that the formulation is able to identify the true values of the parameters under consideration, given that enough information is available.

LP has the consistency property. Indeed, as new alternatives are added in an existing reference set and given that these alternatives add new information (i.e., they are not dominated by alternatives already belonging in the reference set), then the new alternatives will add new non-redundant constraints in LP. These constraints reduce the size of the feasible set. Asymptotically, for large reference sets, this will lead to the identification of a unique optimal solution that represents the decision-maker's judgment policy and preferential system.

2.2 The Multigroup Hierarchical Discrimination Method (MHDIS)

2.2.1 Outline and Main Characteristics

People often employ, sometimes intuitively, a sequential/hierarchical process to classify alternatives to groups using available information and holistic judgments. For example, examine if an alternative can be assigned to the best group C_1, if not then try the second-best group C_2, etc. This is the logic of the MHDIS method and (Zopounidis and Doumpos, 2000b) its main distinctive feature compared with the UTADIS method. A second major difference between the two methods involves the mathematical programming framework used to develop the classification models.

Model development in UTADIS is based on a linear programming formulation followed by a post-optimality stage. In MHDIS, the model development process is performed using two linear programs and a mixed integer one that gradually calibrate the developed model so that it accommodates two objectives: (1) the minimization of the total number of misclassifications, and (2) the maximization of the clarity of the classification. These two objectives are pursued through a lexicographic approach, i.e., initially the minimization of the total number of misclassifications is sought and then the maximization of the clarity of the classification is performed. The common feature shared by both MHDIS and UTADIS involves the form of the criteria aggregation model that is used to model the decision-maker's preferences in classification problems, i.e., both methods employ a utility-based framework.

2.2.2 The Hierarchical Discrimination Process

The MHDIS method proceeds progressively in the classification of the alternatives into the predefined groups. The hierarchical discrimination process used in MHDIS consists of $q-1$ stages (Figure 2.6). Each stage k is considered as a two-group classification problem, where the objective is to discriminate the alternatives of group C_k from the alternatives of the other groups. Because the groups are defined in an ordinal way, this is translated to the discrimination of group C_k from the set of groups $\{C_{k+1}, C_{k+2}, \ldots, C_q\}$. Therefore at each stage of the hierarchical discrimination process, two choices are available for the classification of an alternative:

1. To decide that the alternative belongs in group C_k, or
2. To decide that the alternative belongs at most in the group C_{k+1} (i.e., it belongs in one of the groups C_{k+1} to C_q).

Within this framework, the procedure starts from group C_1 (most preferred alternatives). The alternatives found to belong in group C_1 (correctly or incorrectly) are excluded from further consideration. In a second stage, the objective is to identify the alternatives belonging in group C_2. Once again, all the alternatives found to belong in this group (correctly or incorrectly) are excluded from further consideration, and the same procedure continues until all alternatives are classified into the predefined groups.

The criteria aggregation model used to decide upon the classification of the alternatives at each stage k of the hierarchical discrimination process has the form of an additive utility function, similar to the one used in UTADIS.

$$U_k(\mathbf{g}) = \sum_{i=1}^{n} u_{ki}(g_i) \in [0,1] \tag{2.32}$$

$U(\mathbf{g})$ denotes the utility of classifying any alternative into group C_k on the basis of the alternative's performance on the set of criteria $\mathbf{g}$, and $u_{ki}(g_i)$ denotes the corresponding marginal utility function regarding the classification of any alternative

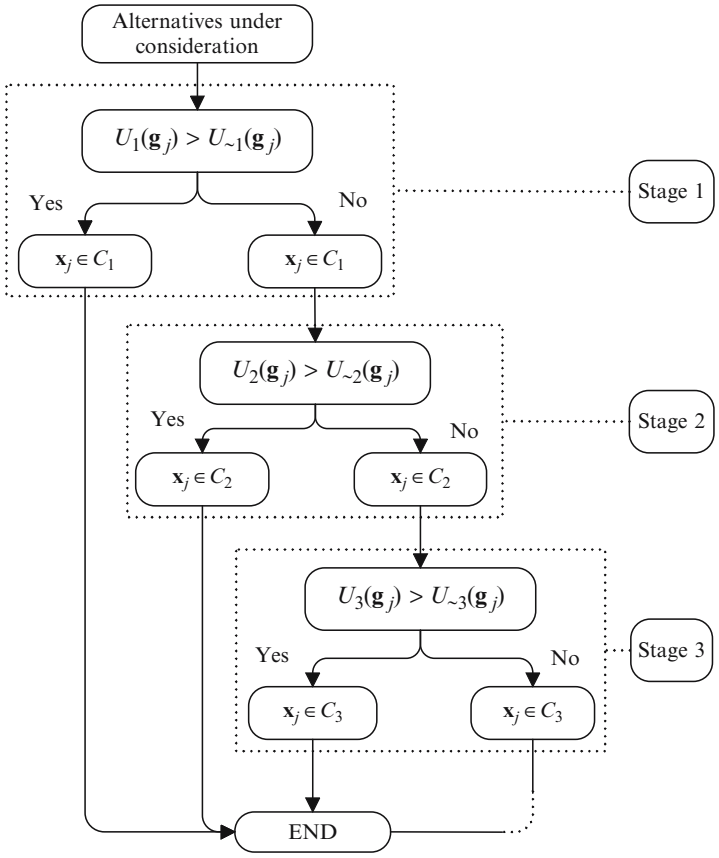

Fig. 2.6 The hierarchical discrimination process in MHDIS

into group C_k according to a specific criterion g_i. Conceptually, the utility function $U_k(\mathbf{g})$ provides a measure of the similarity of the alternatives to the characteristics of group C_k.

Nevertheless, as noted above, at each stage k of the hierarchical discrimination process there are two choices available for the classification of an alternative, the classification into group C_k and the classification at most into group C_{k+1}. The utility function $U_k(\mathbf{g})$ measures the utility (value) of the first choice. To make a classification decision, the utility of the second choice (i.e., classification at most into group C_{k+1}) needs also to be considered. This is measured by a second utility function denoted by $U_{\sim k}(\mathbf{g})$ that has the same form (2.32).

Based on these two utility functions, the classification of an alternative $\mathbf{x}_j$ is performed using the following rules:

$$\left.\begin{array}{l} \text{if } U_k(\mathbf{g}_j) > U_{\sim k}(\mathbf{g}_j) \text{ then } \mathbf{x}_j \in C_k \\ \text{if } U_k(\mathbf{g}_j) < U_{\sim k}(\mathbf{g}_j) \text{ then } \mathbf{x}_j \in C_k^{>} \end{array}\right\} \tag{2.33}$$

where $C_k^{>}$ denotes the set of groups $\{C_{k+1}, C_{k+2}, \ldots, C_q\}$. During model development, the case $U_k(\mathbf{g}_j) = U_{\sim k}(\mathbf{g}_j)$ is considered to be a misclassification. When the developed additive utility functions are used for extrapolating purposes, such a case indicates that the classification of the alternatives is not clear and additional analysis is required. This analysis can be based on the examination of the marginal utilities $u_{ki}(g_{ji})$ and $u_{\sim ki}(g_{ji})$ to determine how the performance of the alternatives on each of the evaluation criterion affects their classification.

In both utility functions $U_k(\mathbf{g})$ and $U_{\sim k}(\mathbf{g})$, the corresponding marginal utilities $u_{ki}(g_{ji})$ and $u_{\sim ki}(g_{ji})$ are monotone functions on the criteria scale. The marginal utility functions $u_{ki}(g_{ji})$ are increasing, whereas $u_{\sim ki}(g_{ji})$ are decreasing functions. This specification is based on the ordinal definition of the groups. In particular, because the alternatives of group C_k are considered to be preferred to the alternatives of the groups C_{k+1} to C_q, it is expected that the higher the performance of an alternative on criterion g_i, the more similar the alternative is to the characteristics of group C_k (increasing form of the marginal utility function $u_{ki}(g_{ji})$) and the less similar it is to the characteristics of the groups C_{k+1} to C_q (decreasing form of the marginal utility function $u_{\sim ki}(g_{ji})$).

The marginal utility functions are modeled in a piece-wise linear form, similar to the case of the UTADIS method. The piece-wise linear modeling of the marginal utility functions in the MHDIS method is illustrated in Figure 2.7. In contrast with the UTADIS method, the criteria's scale is not divided into subintervals. Instead, the performance of each reference alternative is considered as a distinct criterion level. For instance, assuming that the reference set includes m alternatives each having a different performance on criterion g_i, then m criterion levels are considered, ordered from the least preferred one $g_{i*} = \min\{g_{ji}\}$, $\forall \mathbf{x}_j \in A$ to the most preferred one $g_i^* = \max\{g_{ji}\}$, $\forall \mathbf{x}_j \in A$, where a_i is the number of unique values for criterion g_i (e.g., in this example $a_i = m$). Denoting as g_i^h and g_i^{h+1} two consecutive levels of criterion g_i ($g_i^{h+1} > g_i^h$), the monotonicity of the marginal utilities is imposed through the following constraints (z is a small positive constant):

$$w_{kih} \geq z \quad \text{and} \quad w_{\sim kih} \geq z$$

where,

$$w_{kih} = u_{ki}(g_i^{h+1}) - u_{ki}(g_i^h)$$
$$w_{\sim kih} = u_{\sim ki}(g_i^h) - u_{\sim ki}(g_i^{h+1})$$

Thus, it is possible to express the global utility of an alternative $\mathbf{x}_j$ in terms of the incremental variables w as follows:

$$U_k(\mathbf{g}_j) = \sum_{i=1}^{n} \sum_{h=1}^{r_{ji}-1} w_{kih} \quad \text{and} \quad U_{\sim k}(\mathbf{g}_j) = \sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kij} \tag{2.34}$$

Although both UTADIS and MHDIS employ a utility-based modeling framework, it should be emphasized that the marginal utility functions in MHDIS do not

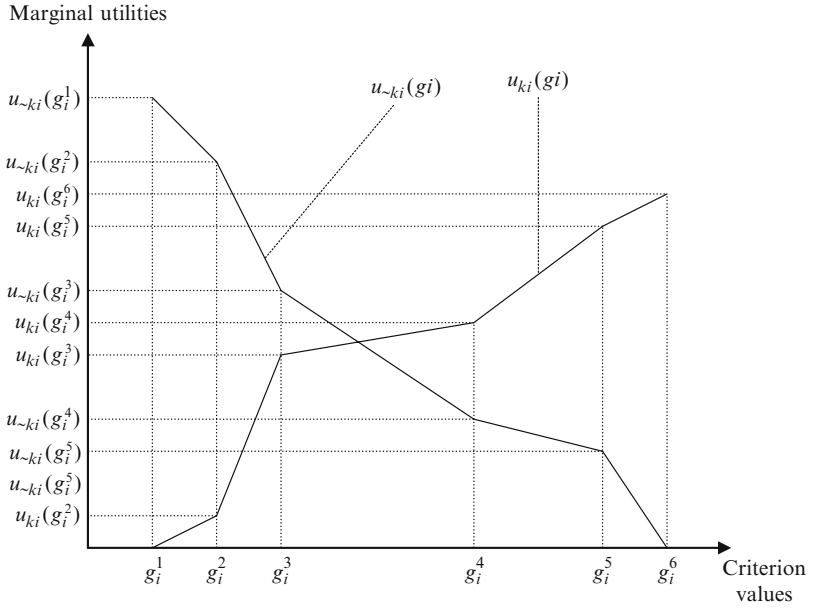

Fig. 2.7 Piece-wise linear form of the marginal utility functions in MHDIS

indicate the performance of an alternative with regard to an evaluation criterion; they rather serve as a measure of the conditional similarity of an alternative $\mathbf{x}_j$ to the characteristics of group C_k (on the basis of a specific criterion) when the choice among C_k and all the lower (worse) groups $C_{k+1},\ldots,C_q$ is considered. In this regard, a high marginal utility $u_{ki}(g_{ji})$ would indicate that when considering the performance of alternative $\mathbf{x}_j$ on criterion g_i, the most appropriate decision would be to assign the alternative into group C_k instead of the set of groups $\{C_{k+1},\ldots,C_q\}$ (the overall classification decision depends upon the examination of all criteria). This simple example indicates that the use of utilities in MHDIS does not correspond to the alternatives themselves, but rather to the appropriateness of the choices (classification decisions) that the decision maker has measured on the basis of the alternatives' performances on the evaluation criteria.

2.2.3 Estimation of Utility Functions

According to the hierarchical discrimination procedure described above, the classification of the alternatives in q classes requires the development of $2(q-1)$ utility functions. The estimation of these utility functions in MHDIS is accomplished through mathematical programming techniques. In particular, at each stage of the hierarchical discrimination procedure, two linear programs and a mixed-integer one

are solved to estimate "optimally" both utility functions.[1] The term "optimally" refers to the classification of the alternatives of the reference set, such that (1) the total number of misclassifications is minimized and (2) the clarity of the classification is maximal.

These two objectives are addressed lexicographically through the sequential solution of two linear programming problems (LP1 and LP2) and a mixed-integer programming problem (MIP). Essentially, the rationale behind the sequential solution of these mathematical programming problems is the following. As noted in the discussion of the UTADIS method, the direct minimization of the total classification error (cf. equations (2.4) or (2.5)) is a quite complex and hard problem to face, from a computational effort point of view. To cope with this problem in UTADIS, an approximation was introduced (cf. equation (2.6)) considering the magnitude of the violations of the classification rules, rather than the number of violations, which defines the classification error rate. As noted, this approximation overcomes the problem involving the computational intensity of optimizing the classification error rate. Nevertheless, the results obtained from this new error function are not necessarily optimal when the classification error rate is considered. To address these issues, MHDIS combines the error function (2.6) with the actual classification error rate. In particular, initially an error function of the form of (2.6) is employed to identify the alternatives of the reference set that are hard to classify correctly (i.e., they are misclassified). This is performed through a linear programming formulation (LP1). Generally, the number of these alternatives is expected to be a small portion of the number of alternatives in the reference set. Then, a more direct error minimization approach is used considering only this reduced set of misclassified alternatives. This approach considers the actual classification error (2.4). The fact that the analysis at this stage focuses only a reduced part of the reference set (i.e., the misclassified alternatives) significantly reduces the computational effort required to minimize the actual classification error function (2.4). The minimization of this error function is performed through a MIP formulation. Finally, given the optimal classification model obtained through the solution of MIP, a linear programming formulation (LP2) is employed to maximize the clarity of the obtained classification without changing the groups into which the alternatives are assigned. The details of this three-step process are described below, along with the mathematical programming formulations used at each step.

LP1: Minimizing the Overall Classification Error

The initial step in the model development process is based on a linear programming formulation. In this formulation, the classification errors are considered as

[1] Henceforth, the discussion focuses on the development of a pair of utility functions at stage k of the hierarchical discrimination process. The first utility function $U_k(\mathbf{g})$ characterizes the alternatives of group C_k, whereas the second utility function $U_{\sim k}(\mathbf{g})$ characterizes the alternatives belonging in the set of groups $\{C_{k+1}, C_{k+2}, \ldots, C_q\}$. The same process applies to all stages $k = 1, 2, \ldots, q-1$ of the hierarchical discrimination process.

real-valued variables, defined similar to the error variables σ^+ and σ^- used in the UTADIS method. In the case of the MHDIS method, these error variables are defined through the classification rule (2.33):

$$\sigma_{kj}^+ = \max\{0, U_{\sim k}(\mathbf{g}_j) - U_k(\mathbf{g}_j)\}, \qquad \forall \mathbf{x}_j \in C_k$$
$$\sigma_{kj}^- = \max\{0, U_k(\mathbf{g}_j) - U_{\sim k}(\mathbf{g}_j)\}, \qquad \forall \mathbf{x}_j \in C_k^>$$

Essentially, the error σ^+ indicates the misclassification of an alternative toward a lower (worst) group compared with the one where it actually belongs, whereas the error σ^- indicates a misclassification toward a higher (better) group. Both errors refer to a specific stage k of the hierarchical model development process.

On the basis of the above considerations, the initial linear program (LP1) to be solved is the following:

$$\min \quad \sum_{k=1}^{q} \left[\frac{\sum_{\forall \mathbf{x}_j \in C_k} \left(\sigma_{kj}^+ + \sigma_{kj}^- \right)}{m_k} \right] \tag{2.35}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \sum_{h=1}^{r_{ji}-1} w_{kih} - \sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} + \sigma_{kj}^+ \geq s, \qquad \forall \mathbf{x}_j \in C_k \tag{2.36}$$

$$\sum_{i=1}^{n} \sum_{j=r_{ji}}^{a_i-1} w_{\sim kih} - \sum_{i=1}^{m} \sum_{j=1}^{r_{ji}-1} w_{kih} + \sigma_{kj}^- \geq s, \qquad \forall \mathbf{x}_j \in C_k^> \tag{2.37}$$

$$w_{kih} \geq z, w_{\sim kih} \geq z \tag{2.38}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{a_i-1} w_{kij} = 1, \sum_{i=1}^{m} \sum_{j=1}^{a_i-1} w_{\sim kij} = 1 \tag{2.39}$$

$$\sigma_{kj}^+, \sigma_{kj}^- \geq 0 \tag{2.40}$$

s, t small positive constants

Constraints (2.36)–(2.37) define the classification error variables σ_{kj}^+ and σ_{kj}^-. These constraints are formulated on the basis of the classification rule (2.33) and the global utility functions (2.34). In the right-hand side of these constraints, a small positive constant s is used to impose the inequalities of the classification rule (2.33). This constant is similar to the constants δ_1 and δ_2 used in the linear programming formulation of the UTADIS method. The set of constraints defined in (2.38) is used to ensure the monotonicity of the marginal utility functions, whereas the set of constraints in (2.39) normalize the global utility to range between 0 and 1.

MIP: Minimizing the number of misclassifications

The solution of LP1 leads to the development of an initial pair of utility functions $U_k(\mathbf{g})$ and $U_{\sim k}(\mathbf{g})$ that discriminate group C_k from the groups C_{k+1} to C_q. These utility functions define a classification of the alternatives in the reference set that is optimal considering the classification error measured in terms of the real-valued variables σ_{kj}^+ and σ_{kj}^-. When the classification error rate is considered, however, these utility functions may lead to suboptimal results. Nevertheless, this initial pair of utility functions enables the identification of the alternatives that can be easily classified correctly and the "hard" alternatives. The "hard" alternatives are the ones misclassified by the pair of utility functions developed through the solution of LP1. Henceforth, the set of alternatives classified correctly by LP1 will be denoted by *COR*, whereas the set of misclassified alternatives will be denoted by *MIS*.

Assuming that the set *MIS* includes at least two alternatives, it is possible to achieve a "rearrangement" of the magnitude of the classification errors σ_{kj}^+ and σ_{kj}^- for the misclassified alternatives (alternatives of *MIS*) that will lead to the reduction of the number of misclassifications. However, as it has already been noted, this requires the introduction of binary 0-1 error variables to MIP model. To avoid the increased computational effort required to solve MIP problems, the MIP formulation used in MHDIS considers only the misclassifications that occur through the solution of LP1, while retaining all the correct classifications. Thus, it becomes apparent that actually, LP1 is an exploratory problem whose output is used as input information to MIP. This reduces significantly the number of binary 0-1 variables, which are associated with each misclassified alternative, thus alleviating the computational effort required to obtain a solution.

While this sequential consideration of LP1 and MIP considerably reduces the computational effort required to minimize the classification error rate, it should be emphasized that the obtained classification model may be near optimal instead of globally optimal. This is due to the fact that MIP inherits the solution of LP1. Therefore, the number of misclassifications attained after solving MIP depends on the optimal solution identified by LP1 (i.e., different optimal solutions of LP1 may lead to different number of misclassifications by MIP). Nevertheless, using LP1 as a preprocessing stage to provide an input to MIP provides an efficient mechanism (in terms of computational effort) to obtain an approximation of the globally minimum number of misclassifications. Formally, MIP is expressed as follows:

$$\min \quad \sum_{k=1}^{q} \left[\frac{\sum_{\forall \mathbf{x}_j \in C_k \cap MIS} \left(E_{kj}^+ + E_{kj}^- \right)}{m_k'} \right] \tag{2.41}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \sum_{j=1}^{r_{ji}-1} w_{kih} - \sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} \geq s, \qquad \forall \mathbf{x}_j \in C_k \cap COR \tag{2.42}$$

$$\sum_{i=1}^{n}\sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} - \sum_{i=1}^{n}\sum_{j=1}^{r_{ji}-1} w_{kih} \geq s, \qquad \forall \mathbf{x}_j \in C_k^{>} \cap COR \tag{2.43}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{r_{ji}-1} w_{kih} - \sum_{i=1}^{n}\sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} + E_{kj}^{+} \geq s \qquad \forall \mathbf{x}_j \in C_k \cap MIS \tag{2.44}$$

$$\sum_{i=1}^{n}\sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} - \sum_{i=1}^{n}\sum_{j=1}^{r_{ji}-1} w_{kih} + E_{kj}^{-} \geq s, \qquad \forall \mathbf{x}_j \in C_k^{>} \cap MIS \tag{2.45}$$

$$w_{kih} \geq z, w_{\sim kih} \geq z \tag{2.46}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{a_i-1} w_{kij} = 1, \sum_{i=1}^{m}\sum_{j=1}^{a_i-1} w_{\sim kij} = 1 \tag{2.47}$$

$$E_{kj}^{+}, E_{kj}^{-} \in \{0,1\} \tag{2.48}$$

s,t small positive constants

Constraints (2.42) and (2.43) are used to ensure that all correct classifications achieved by solving LP1 are retained. Constraints (2.44)–(2.45) are used only for the alternatives that were misclassified by LP1 (set *MIS*). Their interpretation is similar to the constraints (2.36) and (2.37) in LP1. Their only difference is the transformation of the real-valued error variables σ^+ and σ^- of LP1 into the binary 0-1 variables E^+ and E^- that indicate the classification status of an alternative. Constraints (2.44)–(2.45) define these binary variables as follows: $E_{kj}^{+} = 1$ indicates that the alternative $\mathbf{x}_j$ of group C_k is classified by the developed model into the set of groups $C_k^{>}$, whereas $E_{kj}^{-} = 1$ indicates that the alternative $\mathbf{x}_j$ belonging in one of the groups C_{k+1} to C_q ($C_k^{>}$) is classified by the developed model into group C_k. Both cases are misclassifications. On the contrary, the cases $E_{kj}^{+} = 0$ and $E_{kj}^{-} = 0$ indicate the correct classification of the alternative $\mathbf{x}_j$. The interpretation of constraints (2.46) and (2.47) has already been discussed for the LP1 formulation. The objective of MIP involves the minimization of a weighted sum of the error variables E^+ and E^-. The weighting is performed considering the number of alternatives in the set *MIS* from each group C_k. This is denoted by m_k'.

LP2: Maximizing the minimum distance

Solving LP1 and then MIP leads to the "optimal" classification of the alternatives, where the term "optimal" refers to the minimization of the number of misclassified alternatives. However, it is possible that the correct classification of some alternatives is "marginal." This situation appears when the classification rules (2.33) are marginally satisfied, i.e., when there is only a slight difference between $U_k(\mathbf{g}_j)$ and $U_{\sim k}(\mathbf{g}_j)$. For instance, assume a pair of utility functions developed such that for an alternative $\mathbf{x}_j$ of group C_k, its global utilities are $U_k(\mathbf{g}_j) = 0.5$ and $U_{\sim k}(\mathbf{g}_j) = 0.498$. Given these utilities and considering the classification rules (2.33), it is obvious that

alternative $\mathbf{x}_j$ is classified in the correct group (i.e., in group C_k). This is, however, a marginal result. Instead, another pair of utility functions for which $U_k(\mathbf{g}_j) = 0.8$ and $U_{\sim k}(\mathbf{g}_j) = 0.1$ is clearly preferred, providing a more clear conclusion.

This issue is addressed in MHDIS through a third mathematical programming formulation used on the basis of the optimal solution of MIP. At this stage the minimum difference d between the global utilities of the correctly classified alternatives identified after solving MIP is introduced:

$$d = \min \left\{ \min_{\mathbf{x}_j \in C_k \cap COR'} \left\{ U_k(\mathbf{g}_j) - U_{\sim k}(\mathbf{g}_j) \right\}, \min_{\mathbf{x}_j \in C_k^{>} \cap COR'} \left\{ U_{\sim k}(\mathbf{g}_j) - U_k(\mathbf{g}_j) \right\} \right\}$$

where COR' denotes the set of alternatives classified correctly by the pair of utility functions developed through the solution of MIP. The objective of this third phase of the model development procedure is to maximize d. This is performed through the following linear programming formulation (LP2).

$$\min \quad d \tag{2.49}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \sum_{j=1}^{r_{ji}-1} w_{kih} - \sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} - d \geq s, \qquad \forall \mathbf{x}_j \in C_k \cap COR' \tag{2.50}$$

$$\sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} - \sum_{i=1}^{n} \sum_{j=1}^{r_{ji}-1} w_{kih} - d \geq s, \qquad \forall \mathbf{x}_j \in C_k^{>} \cap COR' \tag{2.51}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{r_{ji}-1} w_{kih} - \sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} \leq 0 \qquad \forall \mathbf{x}_j \in C_k \cap MIS' \tag{2.52}$$

$$\sum_{i=1}^{n} \sum_{h=r_{ji}}^{a_i-1} w_{\sim kih} - \sum_{i=1}^{n} \sum_{j=1}^{r_{ji}-1} w_{kih} \leq 0, \qquad \forall \mathbf{x}_j \in C_k^{>} \cap MIS' \tag{2.53}$$

$$w_{kih} \geq z, w_{\sim kih} \geq z \tag{2.54}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{a_i-1} w_{kij} = 1, \sum_{i=1}^{m} \sum_{j=1}^{a_i-1} w_{\sim kij} = 1 \tag{2.55}$$

$$d \geq 0 \tag{2.56}$$

s, t small positive constants

Constraints (2.50)–(2.51) involve only the correctly classified alternatives. In these constraints, d represents the minimum absolute difference between the global utilities of each alternative according to the two utility functions. Constraints (2.52)–(2.53) involve the alternatives misclassified after the solution of MIP (set MIS'), and it is used to ensure that they will be retained as misclassified.

After the solution of LP1, MIP, and LP2 at stage k of the hierarchical discrimination process, the "optimal" classification is achieved between the alternatives belonging in group C_k and the alternatives belonging in the groups $C_k^{>}$. The term

"optimal" refers to the number of misclassifications and to the clarity of the obtained discrimination. If the current stage k is the last stage of the hierarchical discrimination process (i.e., $k = q - 1$), then the model development procedure stops because all utility functions required to classify the alternatives have been estimated. Otherwise, the procedure proceeds to stage $k+1$, in order to discriminate between the alternatives belonging in group C_{k+1} and the alternatives belonging in the lower groups $C_{k+1}^{>}$. In stage $k+1$, all alternatives classified by the pair of utility functions developed at stage k into group C_k are not considered. Consequently, a new reference set A' is formed, including all alternatives that remain unclassified in a specific group (i.e., the alternatives classified in stage k in the set of groups $C_k^{>}$).

2.2.4 Model Extrapolation

The classification of a new alternative $\mathbf{x}_j \notin A'$ is performed by descending the hierarchy of Figure 2.6. Initially, the two first additive utility functions $U_1(\mathbf{g})$ and $U_{\sim 1}(\mathbf{g})$ are used to determine whether the new alternative belongs in group C_1 or not. If $U_1(\mathbf{g}_j) > U_{\sim 1}(\mathbf{g}_j)$, then $\mathbf{x}_j \in C_1$ and the procedure stops, and if $U_1(\mathbf{g}_j) < U_{\sim 1}(\mathbf{g}_j)$, then $\mathbf{x}_j \in C_1^{>}$ and the procedure proceeds with the consideration of the next pair of utility functions $U_2(\mathbf{g})$ and $U_{\sim 2}(\mathbf{g})$. If $U_2(\mathbf{g}_j) > U_{\sim 2}(\mathbf{g}_j)$, then $\mathbf{x}_j \in C_2$ and the procedure stops, and if $U_2(\mathbf{g}_j) < U_{\sim 2}(\mathbf{g}_j)$, then $\mathbf{x}_j \in C_2^{>}$ and the procedure continues in the same way until the classification of the new alternative is achieved.

2.3 Statistical and Econometric Techniques

Statistics is the oldest science involved with the analysis of given samples in order to make inferences about an unknown population. The classification problem is addressed by statistical and econometric techniques within this context. These techniques include both univariate and multivariate methods. The former involve the development and implementation of univariate statistical tests that are mainly of descriptive character. For these reasons, such techniques will not be considered in this review. The foundations of multivariate techniques can be traced back to the work of Fisher (1936) on the linear discriminant analysis (LDA). LDA has been the most extensively used methodology for developing classification models for several decades. Approximately a decade after the publication of Fisher's paper, Smith (1947) extended LDA to the more general quadratic form (quadratic discriminant analysis; QDA).

During the subsequent decades, the focus of the conducted research moved toward the development of econometric techniques. The most well-known methods from this field include the linear probability model, logistic regression and probit models. These three methods are actually special forms of regression analysis in cases where the dependent variable is discrete. The linear probability model is only

suitable for two-group classification problems, whereas both logit and probit models are applicable to multi-group problems, too. The latter two methodologies have several significant advantages over discriminant analysis. This has been one of the main reasons for their extensive use.

Despite the criticism on the use of these traditional statistical and econometric approaches, they still remain quite popular both as research tools as well as for practical purposes. This popularity is supported by the existence of a plethora of statistical and econometric software, which contribute to the easy use of these approaches. Furthermore, statistical and econometric techniques are quite often considered in comparative studies investigating the performance of new classification techniques being developed. In this regard, statistical and econometric techniques often serve as a reference point (benchmark) in conducting such comparisons. It is also important to note that under specific data conditions, statistical techniques yield the optimal classification rule.

2.3.1 Discriminant Analysis

Discriminant analysis has been the first multivariate statistical classification method used for decades by researchers and practitioners in developing classification models. In its linear form it was developed by Fisher (1936). Given a training sample consisting of m alternatives whose classification is a priori known, the objective of the method is to develop a set of discriminant functions maximizing the ratio of among-groups to within-groups variance. In the general case where the classification involves q groups, $q-1$ linear functions of the following form are developed:

$$Z_{kl} = a_{kl} + b_{kl1}g_1 + b_{kl2}g_2 + \cdots + b_{kln}g_n$$

where $g_1, g_2, \ldots, g_n$ are the attributes describing the alternatives $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, a_{kl} is a constant term, and $b_{kl1}, b_{kl2}, \ldots, b_{kln}$ are the attributes' coefficients in the discriminant function. The indices k and l refer to a pair of groups C_k and C_l.

The estimation of the model's parameters involves the estimation of the constant terms a_{kl} and the vectors $\mathbf{b}_{kl} = (b_{kl1}, b_{kl2}, \ldots, b_{kln})$. The estimation procedure is based on two major assumptions: (a) the data follow the multivariate normal distribution, and (b) the variance-covariance matrices for each group are equal. Given these assumptions, the estimation of the constant terms and the attributes' coefficients is performed as follows:

$$\begin{aligned} \mathbf{b}_{kl} &= \mathbf{S}^{-1}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_l) \\ a_{kl} &= -(\bar{\mathbf{x}}_k + \bar{\mathbf{x}}_l)'\mathbf{b}_{kl}/2 \end{aligned}$$

where:

- $\bar{\mathbf{x}}_k$ is a $n \times 1$ vector consisting of the attributes' mean values for group C_k,
- $\mathbf{S}$ is the within-groups variance-covariance matrix, defined as follows:

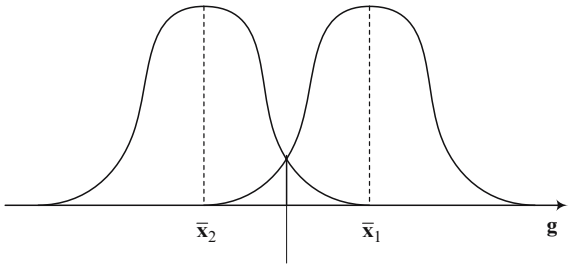

Fig. 2.8 The classification rule in linear discriminant analysis (Source: Altman et al., 1981)

$$\mathbf{S} = \frac{\sum_{k=1}^{q} \sum_{\forall \mathbf{x}_j \in C_k} (\mathbf{x}_j - \bar{\mathbf{x}}_k)(\mathbf{x}_j - \bar{\mathbf{x}}_k)'}{m-q}$$

Once the parameters (coefficients and constant term) of the discriminant functions are estimated, the classification of an alternative $\mathbf{x}_j$ is decided on the basis of its discriminant score $Z_{kl}(\mathbf{x}_j)$ assigned to the alternative by each discriminant function Z_{kl}. In particular, $\mathbf{x}_j$ is classified into group C_k if for all other groups C_l the following rule holds:

$$Z_{kl}(\mathbf{x}_j) \geq \ln \frac{K(k|l)\pi_l}{K(l|k)\pi_k}$$

In the above rule $K(k|l)$ denotes the misclassification cost corresponding to an incorrect decision to classify an alternative into group C_k while actually belong into group C_l and π_k denotes the a priori probability that an alternative belongs into group C_k. Figure 2.8 gives a graphical representation of the above linear classification rule in the two-group case, assuming that all misclassification costs and a priori probabilities are equal.

In the case where the group variance-covariance matrices are not equal, then QDA is used instead of LDA. The general form of the quadratic discriminant function developed through QDA for each pair of groups C_k and C_l is the following:

$$Z_{kl} = a_{kl} + \sum_{i=1}^{n} b_{kli} g_i + \sum_{i=1}^{n} \sum_{h=1}^{n} c_{klih} g_i g_h$$

The estimation of the coefficients and the constant term is performed as follows:

$$\mathbf{b}_{kl} = -2(\bar{\mathbf{x}}'_k \mathbf{S}_k^{-1} - \bar{\mathbf{x}}'_l \mathbf{S}_l^{-1})$$
$$\mathbf{c}_{kl} = \mathbf{S}_k^{-1} - \mathbf{S}_l^{-1}$$
$$a_{kl} = \bar{\mathbf{x}}'_k \mathbf{S}_k^{-1} \bar{\mathbf{x}}_k - \bar{\mathbf{x}}'_l \mathbf{S}_l^{-1} \bar{\mathbf{x}}_l - \ln \left| \mathbf{S}_l \mathbf{S}_k^{-1} \right|$$

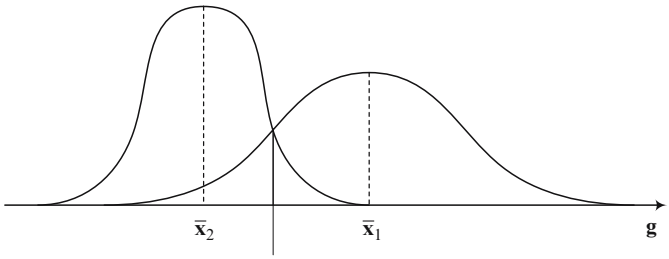

Fig. 2.9 The classification rule in quadratic discriminant analysis (Source: Altman et al., 1981)

$\mathbf{S}_k$ and $\mathbf{S}_l$ denote the within-group variance covariance matrices for groups C_k and C_l, estimated as follows:

$$\mathbf{S}_k = \frac{\sum\limits_{\forall \mathbf{x}_j \in C_k} (\mathbf{x}_j - \bar{\mathbf{x}}_k)(\mathbf{x}_j - \bar{\mathbf{x}}_k)'}{m_k - 1}$$

where m_k is the number of alternatives of the training sample that belong in group C_k.

Given the discriminant score $Z_{kl}(\mathbf{x}_j)$ of an alternative $\mathbf{x}_j$ on every discriminant function corresponding with a pair of groups C_k and C_l, the quadratic classification rule (Figure 2.9) is similar to the linear case: the alternative $\mathbf{x}_j$ is classified into group C_k if and only if for all other groups C_l the following inequality holds:

$$Z_{kl}(\mathbf{g}_j) \geq -2 \ln \frac{K(k|l)\pi_l}{K(l|k)\pi_k}$$

LDA and QDA have been heavily criticized for their underlying assumptions (multivariate normality, known structure of the group variance-covariance matrices). A comprehensive discussion of the impact that these assumptions have on the obtained discriminant analysis' results is presented in the book of Altman et al. (1981).

Given that the above two major underlying assumptions are valid (multivariate normality and known structure of the group variance-covariance matrices), the use of the Bayes rule indicates that the two forms of discriminant analysis (linear and quadratic) yield the optimal classification rule (the LDA in the case of equal group variance-covariance matrices and the QDA in the opposite case). In particular, the developed classification rules are asymptotically optimal (as the training sample size increases, the statistical properties of the considered groups approximate the unknown properties of the corresponding populations). A formal proof of this finding is presented by Duda and Hart (1978), as well as by Patuwo et al. (1993).

Such restrictive statistical assumptions, however, are rarely met in practice. This fact raises a major issue regarding the real effectiveness of discriminant analysis

in realistic conditions. Several studies have addressed this issue. Moore (1973), Krzanowski (1975, 1977), and Dillon and Goldstein (1978) showed that when the data include discrete variables, then the performance of discriminant analysis deteriorates especially when the attributes are significantly correlated. On the contrary, Lanchenbruch et al. (1973) and Subrahmaniam and Chinganda (1978) concluded that even in the case of non-normal data, the classification results of discriminant analysis models are quite robust, especially in the case of the QDA and for data with small degree of skewness.

2.3.2 Logit and Probit Analysis

The aforementioned problems regarding the assumptions made by discriminant analysis motivated researchers to develop more flexible methodologies. The first of such methodologies to be developed includes the linear probability model, as well as logit and probit analysis.

The linear probability model is based on a multivariate regression using as dependent variable the classification of the alternatives of the training sample. Theoretically, the result of the developed model is interpreted as the probability that an alternative belongs in one of the prespecified groups. Performing the regression, however, does not ensure that the model's result lies in the interval [0, 1], thus posing a major model interpretation problem, which makes the use of the linear probability model cumbersome, both from a theoretical and a practical perspective. For these reasons, the use of the linear probability model is rather limited and consequently it will not be further considered in this book.

Logit and probit analysis originate from the field of econometrics. Both models are based on the development of a non-linear function measuring the group-membership probability for the alternatives under consideration. The difference between the two approaches involves the form of the function that is employed. In particular, logit analysis employs the logistic function, whereas the cumulative probability density function of the normal distribution is used in probit analysis. On the basis of these functions, and assuming a two-group classification problem, the probability that an alternative $\mathbf{x}_j$ belongs in group C_2 is defined as follows[2]:

$$\text{Logit analysis: } P_j = F(a+\mathbf{g}_j'\mathbf{b})\frac{1}{1+e^{-a-\mathbf{g}'_j\mathbf{b}}} \tag{2.57}$$

$$\text{Probit analysis: } P_j = f(a+\mathbf{g}_j'\mathbf{b})\int_{-\infty}^{a+\mathbf{g}'_j\mathbf{b}}\frac{1}{(2\pi)^{1/2}}e^{\frac{-z^2}{2}}dz \tag{2.58}$$

[2] If a binary 0-1 coding is used to designate each group such that $C_1 \to 0$ and $C_2 \to 1$, then equations (2.57)–(2.58) provide the probability that an alternative belongs in group C_2. If the binary coding is applied in the opposite way (i.e., $C_1 \to 1$ and $C_2 \to 0$), then equations (2.57)–(2.58) provide the probability that an alternative belongs in group C_1.

Table 2.1 The ordered logit and probit models

Ordered logit model	$P_{1j} = F\left(a_1 + \mathbf{g}'_j\mathbf{b}\right)$
	$P_{2j} = F\left(a_2 + \mathbf{g}'_j\mathbf{b}\right) - F\left(a_1 + \mathbf{g}'_j\mathbf{b}\right)$
	.....
	$P_{kj} = 1 - \left(P_{1j} + P_{2j} + \cdots + P_{k-1,j}\right)$
Ordered probit model	$P_{1j} = \int_{-\infty}^{a_1+\mathbf{g}'_j\mathbf{b}} f(z)dz$
	$P_{2j} = \int_{a_1+\mathbf{g}'_j\mathbf{b}}^{a_2+\mathbf{g}'_j\mathbf{b}} f(z)dz$
	.....
	$P_{kj} = \int_{a_{k-1}+\mathbf{g}'_j\mathbf{b}}^{+\infty} f(z)dz$

The estimation of the constant term a and the vector $\mathbf{b}$ is performed using maximum likelihood techniques. In particular, the parameters' estimation process involves the maximization of the following likelihood function:

$$\ln L = \sum_{\forall \mathbf{x}_j \in C_2} \ln(P_j) + \sum_{\forall \mathbf{x}_j \in C_1} \ln(1 - P_j)$$

The maximization of this function is a nonlinear optimization problem. Altman et al. (1981) report that if there exists a linear combination of the attributes $g_1, g_2, \ldots, g_n$ that accurately discriminates the prespecified groups, then the optimization process will not converge to an optimal solution.

Once the parameters' estimation process is completed, equations (2.57) and (2.58) are used to estimate the group-membership probabilities for all the alternatives under consideration. The classification decision is taken on the basis of these probabilities. For instance, in a two-group classification problem, one can impose a classification rule of the following form: "assign an alternative to group C_2 if $P_j \geq 0.5$; otherwise assign the alternative into group C_1." Alternate probability cutoff points, other than 0.5, can also be specified through trial and error processes.

In the case of multigroup classification problems, logit and probit analysis can be used in two forms: as multinomial or ordered logit/probit models. The difference among multinomial and ordered models is that the former assume a nominal definition of the groups, whereas the latter assume an ordinal definition. In this respect, ordered models are more suitable for addressing sorting problems, and traditional discrimination/classification problems are addressed through multinomial models.

The ordered models require the estimation of a vector of attributes' coefficients $\mathbf{b}$ and a vector of constant terms $\mathbf{a}$. These parameters are used to specify the probability P_{kj} that an alternative $\mathbf{x}_j$ belongs in group C_k, in the way presented in Table 2.1, where $f(z)$ is the standard normal density function.

The constant terms are defined such that $a_{k-1} > a_{k-2} > \cdots > a_2 > 0$ ($a_1 = 0$). The parameters' estimation process is performed similar to the two-group case using maximum likelihood techniques.

The multinomial models require the estimation of a set of coefficient vectors $\mathbf{b}_k$ and a constant term a_k corresponding with each group C_k ($k = 1, 2, \ldots, q$). On the

basis of these parameters, the multinomial logit model estimates the probability P_{kj} that an alternative $\mathbf{x}_j$ belongs in group C_k as follows:

$$P_{kj} = \frac{e^{\mathbf{g}'_j \mathbf{b}_k + a_k}}{\sum_{l=1}^{q} e^{\mathbf{g}'_j \mathbf{b}_k + a_l}}$$

For normalization purposes, $\mathbf{b}_1$ and a_1 are set such that $\mathbf{b}_1 = \mathbf{0}$ and $a_1 = 0$, whereas all other $\mathbf{b}_k$ and a_k $(k = 2, \ldots, q)$ are estimated through maximum likelihood techniques.

Between the logit and probit models, the former is usually preferred. This is mainly because the development of logit models requires less computational effort. Furthermore, there are not strong theoretical and practical results to support a comparative advantage of probit models in terms of their classification accuracy.

During the past three decades, both logit and probit analyses have been extensively used by researchers in a wide range of fields as efficient alternatives to discriminant analysis. However, despite the theoretical advantages of these approaches over LDA and QDA (logit and probit analyses do not pose assumptions on the statistical distribution of the data or the structure of the group variance-covariance matrices), comparative studies made have not clearly shown that these techniques outperform discriminant analysis (linear or quadratic) in terms of their classification performance (Krzanowski, 1975; Press and Wilson, 1978).

2.4 Non-parametric Techniques

In practice, the statistical properties of the data are rarely known, because the underlying population is difficult to be fully specified. This poses problems on the use of statistical techniques and motivated researchers toward the development of non-parametric methods. Such approaches have no underlying statistical assumptions and consequently it is expected that they are flexible enough to adjust to the characteristics of the data under consideration. In the subsequent sections, the most important of these techniques are described.

2.4.1 Neural Networks

Neural networks, often referred to as artificial neural networks, have been developed by artificial intelligence researchers as an innovative modeling methodology of complex problems. The foundations of the neural networks paradigm lie on the emulation of the operation of the human brain. The human brain consists of a huge number of neurons organized in a highly complex network. Each neuron is an individual processing unit. A neuron receives an input signal (stimulus from body

sensors or output signal from other neurons), which after a processing phase produces an output signal that is transferred to other neurons for further processing. The result of the overall process is the action or decision taken in accordance with the initial stimulus.

This complex biological operation constitutes the basis for the development of neural network models. Every neural network is a network of parallel processing units (neurons) organized into layers. A typical structure of a neural network (Figure 2.10) includes the following structural elements:

1. An input layer consisting of a set of nodes (processing units-neurons) one for each input to the network.
2. An output layer consisting of one or more nodes depending on the form of the desired output of the network. In classification problems, the number of nodes of the output layer is determined depending on the number of groups. For instance, for a two-group classification problem the output layer may include only one node taking two values: 1 for group C_1 and 0 for group C_2 (these are arbitrary chosen values and any other pair is possible). In the general case where there are q groups, the number of output nodes is set equal to the number of groups.
3. A series of intermediate layers referred to as hidden layers. The nodes of each hidden layer are fully connected with the nodes of the subsequent and the proceeding layer. Furthermore, it is also possible to consider more complicated structures where all layers are fully connected to each other. Such general network structures are known as fully connected neural networks. The network presented in Figure 2.10 is an example of such structure. There is no general rule to define the number of hidden layers. This is, usually, performed through trial and error processes. Recently, however, a significant part of the research has been devoted to the development of self-organizing neural network models, that is neural networks that adjust their structure to best match the given data conditions. Research made on the use of neural networks for classification purposes showed that, in many cases, a single hidden layer is adequate (Patuwo et al., 1993; Subramanian et al., 1993).

Each connection between two nodes of the network is assigned a weight representing the strength of the connection. The determination of these weights (training of the network) is accomplished through optimization techniques. The objective of the optimization process is to minimize the differences between the recommendations of the network and the actual classification of the alternatives belonging in the training sample.

The most widely used network training methodology is the back propagation approach (Rumerlhart et al., 1986). Recently, advanced nonlinear optimization techniques have also contributed to obtaining globally optimum estimations of the network's connection weights (Hung and Denton, 1993).

On the basis of the connections' weights, the input to each node is determined as the weighted average of the outputs of all other nodes with which there are established connections. In the general case of a fully connected neural network (cf. Figure 2.10) the input in_{ir} to node i of the hidden layer r is defined as follows:

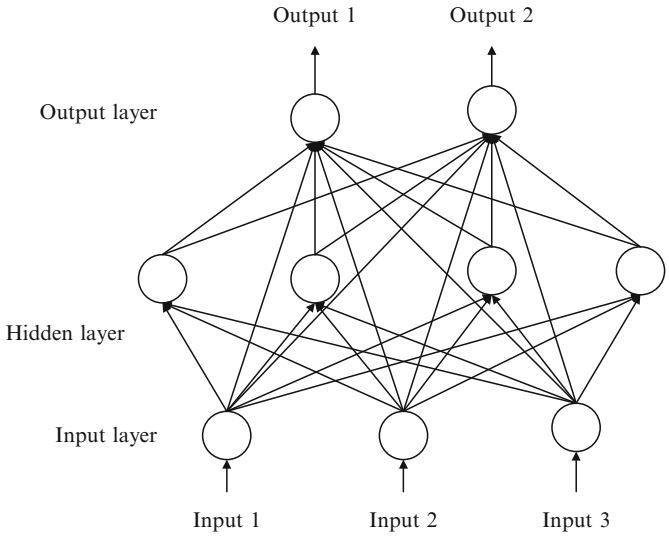

Fig. 2.10 A general architecture of a neural network

$$in_{ir} = \sum_{j=0}^{r-1} \sum_{k=1}^{n_j} w_{ik}^{j} o_{kj} + \phi_{ir}$$

where:

n_j the number of nodes at the hidden layer j,

w_{ik} the weight of the connection between node i of layer r and node k of layer j,

o_{kj} the output of node k at layer j,

ϕ_{ir} an error term.

The output of each node is specified through a transformation function. The most common form of this function is the logistic function:

$$o_{ir} = \frac{1}{1 + e^{-\frac{in_{ir}}{T}}}$$

where T is a user-defined constant.

The major advantage of neural networks is their parallel processing ability as well as their ability to represent highly complex, nonlinear systems. Theoretically, this enables the approximation of any real function with infinite accuracy (Kosko, 1992). These advantages led to the widespread application of neural networks in many research fields. On the other hand, the criticism of the use of neural networks is focused on two points:

1. The increased computational effort required for training the network (specification of connections' weights).

2. The inability to provide explanations of the network's results. This is a significant shortcoming, mainly from a decision support perspective, as in a decision-making context, the justification of the final decision is often a crucial point.

Except for the above two problems, research studies investigating the classification performance of neural networks as opposed to statistical and econometric techniques have led to conflicting results. Subramanian et al. (1993) compared neural networks with LDA and QDA through a simulation experiment using data conditions that were in accordance with the assumptions of the two statistical techniques. Their results show that neural networks can be a promising approach, especially in cases of complex classification problems involving more than two groups and a large set of attributes. On the other hand, LDA and QDA performed better when the sample size was increased.

A similar experimental study by Patuwo et al. (1993) leads to the conclusion that there are many cases where statistical techniques outperform neural networks. In particular, the authors compared neural networks with LDA and QDA, considering both the case where the data conditions are in line with the assumptions of these statistical techniques, as well as the opposite case. According to the obtained results, when the data are multivariate normal with equal group variance-covariance matrices, then LDA outperforms neural networks. Similarly in the case of multivariate normality with unequal variance-covariance matrices, QDA outperformed neural networks. Even in the case of non-normal data, the results of the analysis did not show any clear superiority of neural networks, at least compared with QDA.

The experimental analysis of Archer and Wang (1993) is also worth mentioning. The authors discussed the way that neural networks can be used to address sorting problems and compared their approach with LDA. The results of this comparison show a higher classification performance for the neural networks approach, especially when there is a significant degree of group overlap.

2.4.2 Rule Induction and Decision Trees

During the past two decades, machine learning evolved as a major discipline within the field of artificial intelligence. Its objective is to describe and analyze the computational procedures required to extract and organize knowledge from the existing experience. Within the different learning paradigms (Kodratoff and Michalski, 1990), inductive learning through examples is the one most widely used.

In contrast with the classification techniques described in the previous sections, inductive learning introduces a completely different approach in modeling the classification problem. In particular, inductive learning approaches organize the extracted knowledge in a set of decision rules of the following general form:

IF *elementary conditions* THEN *conclusion*

The first part of such rules examines the necessary and sufficient conditions required for the conclusion part to be valid. The elementary conditions are connected

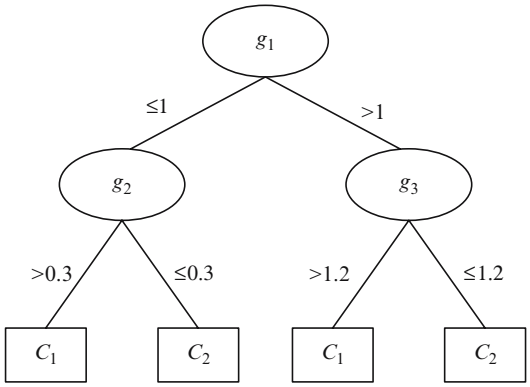

Fig. 2.11 A sample classification decision tree developed using the C4.5 algorithm

using the AND operator. The conclusion consists of a recommendation on the classification of the alternatives satisfying the conditions part of the rule.

One of the most widely used techniques developed on the basis of the inductive learning paradigm is the C4.5 algorithm (Quinlan, 1993). The decision rules developed through the C4.5 algorithm are organized in the form of a decision tree such as the one presented in Figure 2.11. Every node of the tree considers an attribute, and the branches correspond with elementary conditions defined on the basis of the node attributes. Finally, the leaves designate the group to which an alternative is assigned, given that it satisfies the branches' conditions.

The development of the classification tree is performed through an iterative process. Every stage of this process consists of three individual steps:

1. Evaluation of the discriminating power of the attributes in classifying the alternatives of the training sample.
2. Selection of the attribute having the highest discriminating power.
3. Definition of subsets of alternatives on the basis of their performances on the selected attribute.

This procedure is repeated for every subset of alternatives formed in the third step, until all alternatives of the training sample are correctly classified. The evaluation of the attributes' discriminating power in the first step of the above process is performed on the basis of the amount of new information introduced by each attribute in the classification of the alternatives.

The entropy of the classification introduced by each attribute is used as the appropriate information measure. In particular, assuming that each attribute introduces a partitioning of the training sample into t subsets $D_1, D_2, \ldots, D_t$, each consisting of v_t alternatives, then the entropy of this partitioning is defined as follows:

$$I(D) = -\sum_{h=1}^{t} \frac{v_h}{m} \sum_{k=1}^{q} p(D_h/C_k) \log_2[p(D_h/C_k)]$$

where, $p(D_h/C_k)$ denotes the number of alternatives of set D_h that belong in group C_k. The attribute with the minimum entropy is selected as the one with the highest discriminating power. This attribute adds the highest amount of new information in the classification of the alternatives.

The above procedure may lead to a highly specialized classification tree with nodes covering only one alternative. This is the result of overfitting the tree to the given data of the training sample, a phenomenon that is often related to poor generalizing performance. C4.5 addresses this problem through the implementation of a pruning phase, so that the decision tree's size is reduced, in order to improve its expected generalizing performance. The development and implementation of pruning methodologies is a significant research topic in the machine learning community. Some characteristic examples of pruning techniques are the ones presented by Breiman et al. (1984), Gelfand et al. (1991), and Quinlan (1993).

The general aspects of the paradigm used in C4.5 are common to other machine learning algorithms. Some well-known examples of such algorithms include CN2 (Clark and Niblett, 1989), the AQ family of algorithms (Michalski, 1969), and the recursive partitioning algorithm (Breiman et al., 1984).

The main advantages of machine learning classification algorithms involve the following capabilities:

1. Handling of qualitative attributes.
2. Flexibility in handling missing information.
3. Exploitation of large data sets for model development purposes through computationally efficient procedures.
4. Development of easily understandable classification models.

2.4.3 Fuzzy Set Theory

Decision making is often based on fuzzy, ambiguous and vague judgments. The daily use of verbal expressions such as "almost," "usually," "often," etc., are simple yet typical examples of this remark. The fuzzy nature of these simple verbal statements is indicative of the fuzziness encountered in the decision-making process. The fuzzy set theory developed by Zadeh (1965) provides the necessary modeling tools for the representation of uncertainty and fuzziness in complex real-world situations.

The core of this innovative approach is the fuzzy set concept. A fuzzy set is a set with no crisp boundaries. In the case of a traditional crisp set a proposition of the form "alternative **x** belongs to the set *A*" is either true or false; for a fuzzy set, however, it can be partly true or false. Within the context of the fuzzy set theory, the modeling of such fuzzy judgments is performed through the definition of membership functions. A membership function defines the membership degree that an object (alternative) belongs in a fuzzy set. The membership degree ranges in the interval [0, 1]. In the aforementioned example, a membership degree equal to 1 indicates that the proposition "alternative **x** belongs to the set *A*" is true. Similarly, if the membership degree is 0, then it is concluded that the proposition is false. Any other

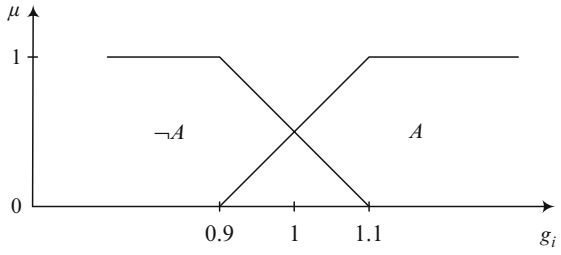

Fig. 2.12 An example of a membership function

value for the membership degree between 0 and 1 indicates that the proposition is partly true.

Figure 2.12 presents an example of a typical form for the membership function for the proposition "according to attribute g_i, alternative $\mathbf{x}$ belongs to the set A." The membership function corresponding with the negation of this proposition is also presented (the negation defines the complement set of A, denoted as $\neg A$).

In order to derive an overall conclusion regarding the membership of an alternative into a fuzzy set based on the consideration of all attributes, one must aggregate the partial membership degrees for each individual attribute. This aggregation is based on common operators such as "AND" and "OR" operators. The former corresponds with the union operation, whereas the latter indicates a intersection operation. A combination of these two operators is also possible.

In the case of classification problems, each group can be considered as a fuzzy set. Similar to the machine learning paradigm, classification models developed through approaches that implement the fuzzy set theory have the form of decision rules. The general form of a fuzzy rule used for classification purposes is the following:

$$\text{IF } (g_{j1} \text{ is } A_{1a}) \wedge (g_{j2} \text{ is } A_{2b}) \wedge \ldots \wedge (g_{jn} \text{ is } A_{nc}) \text{ THEN } \mathbf{x}_j \in C_k$$

where each $A_{i(\cdot)}$ corresponds with a fuzzy set defined on the scale of attribute g_i. The strength of each individual condition is defined by the membership degree of the corresponding proposition "according to attribute g_j alternative $\mathbf{x}_j$ belongs to the set $A_{i(\cdot)}$." The rules of the above general form are usually associated with a certainty coefficient indicating the certainty about the validity of the conclusion part.

Procedures for the development of fuzzy rules in classification problems have been proposed by several researchers. Some indicative studies on this field are the ones of Ishibuchi et al. (1992, 1993), Inuiguchi et al. (2000), Bastian (2000), and Oh and Pedrycz (2000).

Despite the existing debate on the relation between the fuzzy set theory and the traditional probability theory, fuzzy sets have been extensively used to address a variety of real-world problems from several fields. Furthermore, several researchers have exploited the underlying concepts of the fuzzy set theory in conjunction with other disciplines such as neural networks (neurofuzzy systems; Von Altrock, 1996), expert systems (fuzzy rule-based expert systems; Langholz et al., 1996),

mathematical programming (fuzzy mathematical programming; Zimmermann, 1978), and MCDA (Yager, 1977; Dubois and Prade, 1979; Siskos, 1982; Siskos et al., 1984; Fodor and Roubens, 1994; Grabisch, 1995, 1996; Lootsma, 1997).

2.4.4 Rough Sets

Pawlak (1982) introduced the rough set theory as a tool to describe dependencies between attributes, to evaluate the significance of attributes, and to deal with inconsistent data. As an approach to handle imperfect data (uncertainty and vagueness), it complements other theories that deal with data uncertainty, such as probability theory, evidence theory, fuzzy set theory, etc. Generally, the rough set approach is a very useful tool in the study of classification problems. Recently, however, there have been several advances in this field to allow the application of the rough set theory to choice and ranking problems as well (Greco et al., 1997).

The rough set philosophy is founded on the assumption that with every alternative, some information (data, knowledge) is associated. This information involves two types of attributes: condition and decision attributes. Condition attributes are those used to describe the characteristics of the objects. For instance, the set of condition attributes describing a country can be a set of economic, political, and social indicators. The decision attributes define a partition of the objects into groups according to the condition attributes.

On the basis of these two types of attributes, an information table $S = \langle U, Q, V, f \rangle$ is formed, as follows:

- U is a finite set of m alternatives (objects).
- Q is a finite set of n attributes.
- V is the intersection of the domains of all attributes (the domain of each attribute g_i is denoted by V_i). The traditional rough set theory assumes that the domain of each attribute is a discrete set. In this context, every quantitative real-valued attribute needs to be discretized[3] using discretization algorithms such as the ones proposed by Fayyad and Irani (1992), Chmielewski and Grzymala-Busse (1996), and Zighed et al. (1998). Recently, however, the traditional rough set approach has been extended so that no discritezation is required for quantitative attributes. Typical examples of the new direction are the DOMLEM algorithm (Greco et al., 1999a) and the MODLEM algorithm (Grzymala-Busse and Stefanowski, 2001).
- $f : U \times Q \rightarrow V$ is a total function such that $f(\mathbf{x}_j, g_i) \in V_i$ for every $g_i \in Q$, $\mathbf{x}_j \in U$, called information function (Pawlak, 1991; Pawlak and Slowinski, 1994).

Simply stated, the information table is an $m \times n$ matrix, with rows corresponding with the alternatives and columns corresponding with the attributes.

Given an information table, the basis of the traditional rough set theory is the indiscernibility between the alternatives. Two alternatives $\mathbf{x}_j$ and $\mathbf{x}_l$ are considered

[3] Discretization involves the partitioning of an attribute's domain $[a, b]$ into h subintervals $[v_1, v_2)$, $[v_2, v_3)$, ..., $[v_{h-1}, v_h]$, where $v_1 = a$ and $v_h = b$.

to be indiscernible, if and only if they are characterized by the same information, i.e., $f(\mathbf{x}_j, g_i) = f(\mathbf{x}_l, g_i)$ for every $g_i \in P \subseteq Q$. In this way, every $P \subseteq Q$ leads to the development of a binary relation on the set of alternatives. This relation is called $P-$indiscernibility relation, denoted by I_P. I_P is an equivalence relation for any P.

Every set of indiscernible alternatives is called elementary set and it constitutes a basic granule of knowledge. Equivalence classes of the relation I_P are called $P-$elementary sets in S, and $I_P(\mathbf{x}_j)$ denotes the $P-$elementary set containing alternative $\mathbf{x}_j \in U$.

Any set of objects being a union of some elementary sets is referred to as crisp (precise) otherwise it is considered to be rough (imprecise, vague). Consequently, each rough set has a boundary line consisting of cases (objects) that cannot be classified with certainty as members of the set or of its complement. Therefore, a pair of crisp sets, called the lower and the upper approximation can represent a rough set. The lower approximation consists of all objects that certainly belong to the set and the upper approximation contains objects that possibly belong to the set. The difference between the upper and the lower approximation defines the doubtful region, which includes all objects that cannot be certainly classified into the set. On the basis of the lower and upper approximations of a rough set, the accuracy of its approximation can be calculated as the ratio of the cardinality of its lower approximation to the cardinality of its upper approximation.

Assuming that $P \subseteq Q$ and $Y \subseteq U$, then the $P-$lower approximation, the $P-$upper approximation, and the $P-$doubtful region of Y ($\underline{P}Y$, $\overline{P}Y$, and $BN_P(Y)$, respectively) are formally defined as follows:

$$\underline{P}Y = \{\mathbf{x}_j \in Y : I_P(\mathbf{x}_j) \in Y\} \tag{2.59}$$

$$\overline{P}Y = \bigcup_{\mathbf{x}_j \in Y} I_p(\mathbf{x}_j) \tag{2.60}$$

$$BN_P(Y) = \overline{P}Y - \underline{P}Y \tag{2.61}$$

On the basis of these approximations, it is possible to estimate the accuracy of the approximation of the rough set Y, denoted by $\alpha_P(Y)$. The accuracy of the approximation is defined as the ratio of the number of alternatives belonging in the lower approximation to the number of alternatives of the upper approximation:

$$\alpha_P(Y) = \frac{|\underline{P}Y|}{|\overline{P}Y|}$$

Within the context of a classification problem, each group C_k is considered as a rough set $_k$. The overall quality of the approximation of the classification by a set of attributes P is defined as follows:

$$\gamma_P(Y) = \frac{\sum_{k=1}^{q} |\underline{P}Y_k|}{m}$$

Having defined the quality of the approximation, the first major capability that the rough set theory provides is to reduce the available information, so as to retain only the information that is absolutely necessary for the description and classification of the alternatives. This is achieved by discovering subsets R of the complete set of attributes P, which can provide the same quality of classification as the whole attributes' set, i.e., $\gamma_P(Y) = \gamma_R(Y)$. Such subsets of attributes are called reducts and are denoted by $\text{RED}_Y(P)$. Generally, the reducts are more than one. In such a case, the intersection of all reducts is called the core, i.e., $\text{CORE}_Y(P) = \cap \text{RED}_Y(P)$. The core is the collection of the most relevant attributes, which cannot be excluded from the analysis without reducing the quality of the obtained description (classification). The decision maker can examine all obtained reducts and proceed to the further analysis of the considered problem according to the reduct that best describes reality. Heuristic procedures can also be used to identify an appropriate reduct (Slowinski and Zopounidis, 1995).

The subsequent steps of the analysis involve the development of a set of rules for the classification of the alternatives into the groups where they actually belong. The rules developed through the rough set approach have the following form:

IF *conjunction of elementary conditions*
THEN *disjunction of elementary decisions*

The procedures used to construct a set of decision rules employ the machine learning paradigm. Such procedures developed within the context of the rough set theory have been presented by Grzymala-Busse (1992), Slowinski and Stefanowski (1992), Skowron (1993), Ziarko et al. (1993), Stefanowski and Vanderpooten (1994), Mienko et al. (1996), and Grzymala-Busse and Stefanowski (2001). Generally, rule induction techniques follow one of the following strategies:

1. Development of a minimal set of rules covering all alternatives of the training sample (information table).
2. Development of an extensive set of rules consisting of all possible decision rules.
3. Development of a set of strong rules, even partly discriminant,[4] which do not necessarily cover all alternatives of the training sample.

Irrespective of the rule induction approach employed, a decision rule developed on the basis of the rough set approach has some interesting properties and features. In particular, if all alternatives that satisfy the condition part belong in the group indicated by the conclusion of the rule, then the rule is called consistent. In the case where the condition part considers only a single group, then the rule is called exact, otherwise the rule is called approximate. The conclusion part of approximate

[4] Rules covering only alternatives that belong to the group indicated by the conclusion of the rule (positive examples) are called discriminant rules. On the contrary, rules that cover both positive and negative examples (alternatives not belonging in the group indicated by the rule) are called partly discriminant rules. Each partly discriminant rule is associated with a coefficient measuring the consistency of the rule. This coefficient is called level of discrimination and is defined as the ratio of positive to negative examples covered by the rule.

rules involves a disjunction of at least two groups ($\mathbf{x}_j \in C_k \vee \mathbf{x}_j \in C_h \vee \ldots$). Approximate rules are developed when the training sample (information table) includes indiscernible alternatives belonging in different groups. Each rule is associated with a strength measure, indicating the number of alternatives covered by the rule. For approximate rules, their strength is estimated for each individual group considered in their conclusion part. Stronger rules consider a limited number of elementary conditions; thus, they are more general.

Once the rule induction process is completed, the developed rules can be easily used to decide upon the classification of any new alternative not considered during model development. This is performed by matching the conditions part of each rule to the characteristics of the alternative, in order to identify a rule that covers the alternative. This matching process may lead to one of the following four situations (Slowinski and Stefanowski, 1994):

1. The alternative is covered only by one exact rule.
2. The alternative is covered by more than one exact rule, all indicating the same classification.
3. The alternative is covered by one approximate rule or by more than one exact rule indicating different classifications.
4. The alternative is not covered by any rule.

The classification decision in situations (1) and (2) is straightforward. In situation (3), the developed rule set leads to conflicting decisions regarding the classification of the alternative. To overcome this problem, one can consider the strength of the rules that cover the alternative (for approximate rule, the strength for each individual group of the condition part must be considered). The stronger rule can be used to take the final classification decision. This approach is employed in the LERS classification system developed by Grzymala-Busse (1992).

Situation (4) is the most difficult one, because using the developed rule set one has no evidence as to the classification of the alternative. The LERS system tackles this problem through the identification of rules that partly cover the characteristics of the alternative under consideration.[5] The strength of these rules as well as the number of elementary conditions satisfied by the alternative are considered in making the decision. An alternative approach proposed by Slowinski (1993) involves the identification of a rule that best matches the characteristics of the alternative under consideration. This is based on the construction of a valued closeness relation measuring the similarity between each rule and the alternative. The construction of this relation is performed in two stages. The first stage involves the identification of the attributes that are in accordance to the affirmation "the alternative $\mathbf{x}_j$ is close to rule r." The strength of this affirmation is measured on a numerical scale between 0 and 1. The second stage involves the identification of the characteristics that are in discordance with the above affirmation. The strength of concordance and discordance tests are combined to estimate an overall index representing the similarity of a rule to the characteristics of the alternative.

[5] Partly covering involves the case where the alternative satisfies only some of the elementary conditions of a rule.

Closing this brief discussion of the rough set approach, it is important to note the recent advances made in this field toward the use of the rough set approach as a methodology of preference modeling in multicriteria decision problems (Greco et al., 1999a, 2000a). The main novelty of the recently developed rough set approach concerns the possibility of handling criteria, i.e., attributes with preference ordered domains, and preference ordered groups in the analysis of sorting examples and the induction of decision rules. The rough approximations of decision groups involve dominance relation, instead of indiscernibility relation considered in the basic rough set approach. They are built of reference alternatives given in the sorting example (training sample). Decision rules derived from these approximations constitute a preference model. Each "if ... then ..." decision rule is composed of (a) a condition part specifying a partial profile on a subset of criteria to which an alternative is compared using the dominance relation, and (b) a decision part suggesting an assignment of the alternative to "at least" or "at most" a given class.

The decision rule preference model has also been considered in terms of conjoint measurement (Greco et al., 2001). A representation theorem for multicriteria sorting proved by Greco et al. states an equivalence of simple cancellation property, a general discriminant (sorting) function, and a specific outranking relation, on the one hand, and the decision rule model on the other hand. It is also shown that the decision rule model resulting from the dominance-based rough set approach has an advantage over the usual functional and relational models because it permits handling inconsistent sorting examples. The inconsistency in sorting examples is not unusual due to instability of preference, incomplete determination of criteria, and hesitation of the decision maker.

It is also worth noting that the dominance-based rough set approach is able to deal with sorting problems involving both criteria and regular attributes whose domains are not preference ordered (Greco et al., 2002) and missing values in the evaluation of reference alternatives (Greco et al., 1999b; Greco et al., 2000b). It also handles ordinal criteria in a more general way than the Sugeno integral, as it has been proved in Greco et al. (2001).

2.5 Miscellaneous Techniques

Apart from the techniques that have been described above, other multi-criteria optimization methodologies could also be used for the evaluation of country risk. Based on Olson and Shi (2005), financial management problems can be data mined using large real-life data sets. Moreover, they support that in the financial business, practitioners have applied a number of data-mining techniques to support credit card portfolio management and further financial management. These techniques include the Behavior Score developed by Fair Isaac Corporation (FICO), Credit Bureau Scores, First Data Resource (FDR)'s Proprietary Bankruptcy Score and Set Enumeration (SE) decision tree.

Kou et al. (2003) promote a multiple criteria linear programming (MCLP) approach to data mining based on linear discriminant analysis. They describe the connections between MCLP and data mining, including several general models of MCLP approaches. Similarly, Kou et al. (2004) propose a classification model by using multiple criteria linear programming to discover behavior patterns of credit card holders. As continuation of this research, He et al. (2004) propose a heuristic classification method by using the fuzzy linear programming to discover the bankruptcy patterns of credit card holders.

Taking into account that credit risk and bankruptcy risk constitute part of country financial risk, the aforementioned methodologies could also be used for the evaluation of country risk.

http://www.springer.com/978-0-387-76679-9

Country Risk Evaluation
Methods and Applications
Kosmidou, K.; Doumpos, M.; Zopounidis, C.
2008, X, 120 p. 17 illus., Hardcover
ISBN: 978-0-387-76679-9