Multicriteria Models for Learning Ordinal Data: a literature review

Ricardo Sousa, Iryna Yevseyeva, Joaquim F. Pinto da Costa, and Jaime S. Cardoso

Abstract Operations Research (OR) and Artificial Intelligence (AI) disciplines have been playing major roles on the design of new intelligent systems. Recently, different contributions from both fields have been made on the models design for problems with multi-criteria. The credit scoring problem is an example of that. In this problem, one evaluates how unlikely a client will default with his payments. Client profiles are evaluated, being their results expressed in terms of an ordinal score scale (*Excelent* $\succ$ *Good* $\succ$ *Fair* $\succ$ *Poor*). Intelligent systems have then to take in consideration different criteria such as payment history, mortgages, wages among others in order to commit their outcome. To achieve this goal, researchers have been delving models capable to render these multiple criteria encompassed on ordinal data.

The literature presents a myriad of different methods either on OR or AI fields for the multi-criteria models. However, a description of ordinal data methods on these two major disciplines and their relations has not been thoroughly conducted yet. It is key for further research to identify the developments made and the present state of the existing methods. It is also important to ascertain current achievements and what the requirements are to attain intelligent systems capable to capture relationships from data. In this chapter one will describe techniques presented for over more than five decades on OR and AI disciplines applied to multi-criteria ordinal problems.

Ricardo Sousa · Jaime S. Cardoso
INESC TEC (formerly INESC Porto), Faculdade de Engenharia, Universidade do Porto
Porto, Portugal
e-mail: {rsousa, jaime.cardoso}@inescporto.pt

Iryna Yevseyeva
Computer Science and Communication Research Center, Polytechnic Institute of Leiria
Leiria, Portugal
e-mail: iryna.yevseyeva@gmail.com

Joaquim F. Pinto da Costa
CMUP, Faculdade de Ciências da Universidade do Porto
Porto, Portugal
e-mail: jpcosta@fc.up.pt

1 Introduction

Learning multicriteria (MC) models from data has recently gathered a substantial attention. Such trend has its reasons in the diverse set of applications which can be found in management [76, 105], financial [31, 33] and medical [6, 125] fields, to name a few. Consequently, the very diversity of the multicriteria learning research topics led to a discussion and proposals in several different fields. Decision analysis, machine learning and statistics/econometrics are some of them. Hence, a rich terminology can be found due to this diverse fields of study. Sorting, ranking, dominance, among others, are some of the many names referring to multicriteria methods. Even though mostly all share the same fundamental principles, it is on the methods assumptions that most differences occur. Nevertheless, methods for learning ordinal data have been recently seen as a generalisation of some multicriteria techniques [2].

The importance of ordinal data is clear. Nowadays, industry tries to cope with current technological advancements and towards profit maximisation. Hence, more and more personalised products and services are being commercialised to a wider audience. Problems like credit scoring where the system evaluates the capability of one default his debts [31, 33, 141] by grading a customer credit profile in the scale *Excelent* $\succ$ *Good* $\succ$ *Fair* $\succ$ *Poor*, movies suggestion [28], breast cancer diagnosis [17], or gene analysis through the analysis of hyperactivity on certain proteins [102, 103], are some examples of ordinal problems where data is structured by a "natural" order. As a result, new and robust computational methods capable to unravel reasoning's behind ones decisions also led to new theoretical developments. Regarding to these developments two major disciplines lead the research: Operations Research (OR) and Artificial Intelligence (AI).

In this chapter it is presented a literature review over different areas that deal with ordinal data, in general, to the extent of what it is known nowadays. Section 2 starts by providing the terminology that will be used. Section 3 will focus on methods on the operations research side whereas techniques in the artificial intelligence field will be described in Section 4. Section 3 will be concerned mainly about aggregation models, fuzzy and rough set theory, and evolutionary algorithms approaches. Section 4 will be dedicated to inductive learning, a very large and important topic within AI. In this section different existing works in the literature will be identified as well as feature selection approaches (Section 4.1) and performance assessment metrics (Section 4.2). As remainder of this chapter, in Section 5, one will draw a summary of what has been achieved until now and what still poses as open problems.

2 Terminology and Concepts

Learning multicriteria (MC) on ordinal data has a strong connection with OR and AI [146]. Albeit being conceptually different topics, there is an intrinsic connection among them. OR comprises several different areas of study such as decision analysis, mathematical programming among others. Whereas, AI can be described as

being composed by machine learning, pattern recognition, data mining [119] etc. Within each area there are concepts borrowed from one another. For instance, machine learning vastly uses techniques from mathematical programming and statistics since its early days [45, 132] (Fig. 1 depicts some of these relations). How these topics interact with each other is not within the scope of this chapter. It is the purpose of Fig. 1 to illustrate the broad aspects of the area in study. Its usage is so broad that a full coverage is not possible. However, it is interesting to know how MC methods
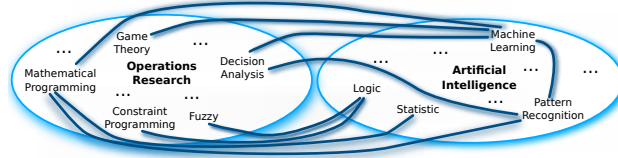

Fig. 1: Illustration of the different fields that overlap with operations research and artificial intelligence.

have been used in data analysis to represent knowledge. Such is done in order to understand reasoning's behind decisions [94], outcome prediction [26, 36], in mimicking behaviours [90] and planning [68, 105].

Even though MC methods have been thoroughly studied, not much effort has been employed on the particular case where data is presented in a "natural" order. Let us consider the credit score problem. A bank assigns a score of *Excellent* to a client given his wage, good payment history in previous mortgages and the number of credits at the time of the evaluation. The score assessment is clearly rendered over the different criteria: Wage, payment history, among others. Ideally, one wants to find the best function that can capture all this information in order to output the expected outcome.

Definition 1 (Classification on Ordinal Data Problems [16, 18, 91, 93, 146]). *Classifying on ordinal data problems consists on finding the best mapping $f : \mathbb{R}^d \to \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ of a given pattern, $\boldsymbol{x} \in \mathbb{R}^d \subset \mathcal{X}$, to a finite set of classes, where $\mathcal{C}_K \succ \ldots \succ \mathcal{C}_1$.*

Pattern $\boldsymbol{x}$ is also referred as instance, example or alternative. Moreover, $\boldsymbol{x}$ can be represented in a vector fashion where each entry is identified as a feature, attribute or criterion, i.e., $\boldsymbol{x} = \{x_1, \ldots, x_d\}$. A dataset is a tuple consisted of N patterns and its target classes (or outcomes), $\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^N$ and $\succ$, the order relation on the set of classes.

Literature usually differentiates attribute and criterion [51]. Consequently, the problem in analysis can be substantially different. In an ordinal data problem as the credit scoring, an alternative (to which client a loan should be granted) is characterised by several criteria, each one representing a level of *importance* to the decision maker (DM) (the bank). Here, criteria is used instead of attribute being the former more adequate for the ordinal problem [51, 135].

The usage of the term ranking is also common in the MC field. However, such term is usually mentioned to other subjects aside classification.

Definition 2 (Ranking [16, 25]). *A ranking problem consists on finding the best mapping $f : \mathbb{R}^d \to \{\mathcal{R}_1, \ldots, \mathcal{R}_L\}$ of a given pattern, $\boldsymbol{x} \in \mathbb{R}^d \subset \mathcal{X}$, to a finite set of ranks, where $\mathcal{R}_L \succ \ldots \succ \mathcal{R}_1$ is* not *pre-defined.*

There are subtle differences between the two problems. Whereas in classification the order between classes is already defined and all patterns have to be assigned into at most one class, in ranking such does not hold. Think for instance on the Google™ or Yahoo™ search engines. When entering a search query, the result can vary from user to user for the same query. The search engine will look on its database and rank the results according to, for instance, user search history. Ranking approaches however go beyond the subject of this chapter.

Depending on the problem, criteria can also represent a magnitude of importance or unimportance, a ratio, among others. This can generate datasets where order may not be explicitly represented. Different works tackled the ordinal problem assuming that data were monotone, i.e., where both criteria and classes were assumed to be ordered [10, 39, 101]. Nevertheless, recent works argue that monotonicity constraint cannot be verified despite being however perfect representatives of an ordinal problem [18, 56]. The following synthetic datasets depict some of those claims. To each
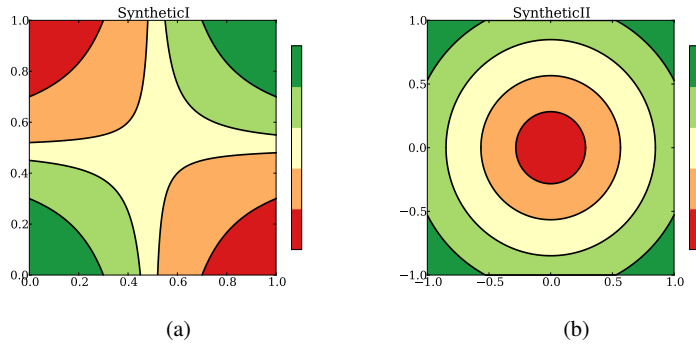

Fig. 2: Two synthetic ordinal dataset where the monotonicity property at input data does not hold.

point in Fig. 2a was assigned a class y from the set $\{1, 2, 3, 4, 5\}$, according to

$$y = \min_{r \in \{1,2,3,4,5\}} \{r : b_{r-1} < 10(x_1 - 0.5)(x_2 - 0.5) + \varepsilon < b_r\}$$

$$(b_0, b_1, b_2, b_3, b_4, b_5) = (-\infty, -1, -0.1, 0.25, 1, +\infty) \tag{1}$$

where $\varepsilon \sim N(0; 0.125^2)$ simulates the possible existence of error in the assignment of the true class to $\boldsymbol{x}$. Data in Fig. 2b is uniformly distributed in the unit-circle, with

the class y being assigned according to the radius of the point: $y = \left\lceil \sqrt{x_1^2 + x_2^2} \right\rceil$. These synthetic datasets are examples where order cannot be captured directly in the input space, but in an implicit feature space.

Hence, the following question can be posed: How to capture order? Many models have been proposed towards this goal. But before answering that question, first a brief description of the most commonly used models is required. The following concepts will allow a better understanding of the most recent techniques discussed along this chapter.

Starting by the OR domain, a classic multicriteria decision analysis (MCDA) approach is done by the representation of a specific aggregation model. Aggregation models are performed by aggregating different value or utility functions in order to be expressed by a single criterion. One aggregation model that we can think of a, for instance, the mean: $\frac{1}{d} \sum_{j=1}^{d} x_j$. The use of utility vs. value depends upon the problem. Whereas, utility functions are used in stochastic problems, value function are used in deterministic ones [92]. In brief, an aggregation model is a function $\mathcal{U} : \mathbb{R}^d \to \mathbb{R}$, that maps criteria of the DM onto outcomes [92]. Utility functions are widely used, where the one presented in Equation (2) is an example of several other aggregation models. It has the advantage of considering both qualitative and quantitative criteria. The simplest additive case of an utility function is defined as follows:

$$\mathcal{U}(\boldsymbol{x}) = \sum_{j=1}^{d} u_j(x_j) \tag{2}$$

where $\mathcal{U} \in [0,1]$. For the interested reader Siskos [115] presents a good description of these methods.

Fuzzy set theory is another topic with increasing interest on the scientific community. Its usage is not restricted only to the MCDA problem being however strongly defended thanks to its capability to handle uncertainty [50, 65]. In general, fuzzy set theory presents a fundamental principle which describes a special type of sets which have *degrees of membership* through simple logical operators. Such can be described by any mapping function $\mu(\boldsymbol{x}) : \mathbb{R}^d \to [0,1]$. Fig. 3a) consists of a valid representation for a given membership function. Moreover, it can represent knowledge in a *if ... then* way in a similar way to decision trees (DTs) [69] which will be described shortly.

In much of the works currently present in the literature, fuzzy set theory usually appears along with rough sets. The latter field is however slightly different from the former. Rough Set theory not just handle uncertainty, but also incomplete information which can be present on data [65]. Even though new approaches on utility additive functions (UTA—UTilitès Additives [115]) already tackle this problem, it has also been stated that rough and fuzzy set theory are complementary because of dealing with different kinds of uncertainty [50]. It was initially proposed by Pawlak [97] with the objective to provide a mathematical formulation of the concept of approximated (rough) equality of sets in a given space. In the rough set theory it is assumed that to every object there is an associated amount of information that describes
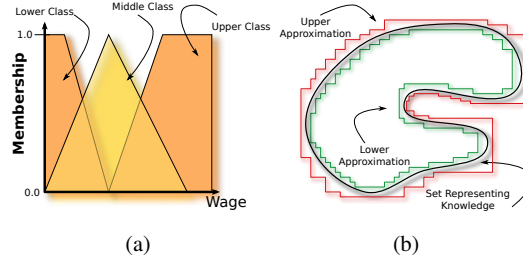
(a)   (b)

Fig. 3: Fuzzy and Rough Set concept illustrations: (a) An example of a membership function that defines a possible economic class problem in a fuzzy set approach; (b) Lower and Upper approximations of a given set which represent the domain knowledge;

it. This refers to the view that knowledge has a granular structure [1, 50, 51, 98]. Therefore, an important characteristic of rough sets theory is the identification of consistent data and assigning them into lower and upper approximations of sets— see Fig. 3b).

More on the AI domain, in general, one tries to obtain valid generalisation rules, classifier, from data. Once a classifier has been designed, one has to assess its performance by estimating the error of the classifier for unseen examples. Classification error is expressed as a misclassification error defined by a "true misclassification rate" (here denoted as $R^*(d)$). $d(\boldsymbol{x})$ is the learner model with input data $\boldsymbol{x}$. Breiman [15] defines this function as:

Definition 3 (Accuracy Estimation [15]). *Take $(\boldsymbol{x}, y)$, $\boldsymbol{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, to be a new sample from the probability distribution $\mathbb{P}(A, j)$; i.e.,*

- *$\mathbb{P}(\boldsymbol{x} \in A, y = j) = \mathbb{P}(A, j)$.*
- *$(\boldsymbol{x}, y)$ is independent of $\mathcal{D}$.*

Then define

$$R^*(d) = \mathbb{P}(d(\boldsymbol{x}) \neq y) \tag{3}$$

But how can $R^*(d)$ be estimated? There are many approaches. One that this work will use is the cross-validation approach. Dataset $\mathcal{D}$ is randomly divided in subsets, with the same size as possible, e.g., $\mathcal{D}_1, \ldots, \mathcal{D}_V$. For each v, $v = 1, \ldots, V$, a learning method is applied to the sample $\mathcal{D} - \mathcal{D}_v$, resulting in the $d^v(\boldsymbol{x})$ model.

$$R^{cv}(d) = \frac{1}{V} \sum_{v=1}^{V} R^{ts}(d^v) \tag{4}$$

where R^{ts} is defined as

$$R^{ts}(d^v) = \frac{1}{N_v} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_v} \mathcal{F}(d^v(\boldsymbol{x}_i), y_i) \tag{5}$$

where $N_v \simeq N/V$ and $\mathcal{F}$ any function which penalises each error[1]. One can now analyse the different learning methods for ordinal data.

k-Nearest Neighbour (k-NN) is a simple method that interestingly has not been explored enough in the MCDA setting until very recently. It consists of a non-parametric method with the main objective to estimate the density function from sample patterns [38]. It extends the local region around a data point $\boldsymbol{x}$ until the kth nearest neighbour is found. The most represented class in the k-closest cases defines the predicted class. Fig. 4a-b) illustrates such procedure. DTs are another
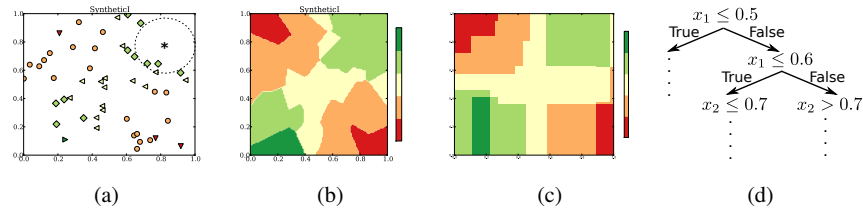

(a) (b) (c) (d)

Fig. 4: k-NN and DT methods. (a) A test pattern (illustrated as a star) composed by two features checks for, in this example, two closest labelled patterns in order to determine its own class; (b) Prediction over the whole feature domain for an 2-NN on the training data shown in (a); (c) A DT discriminates the feature space by rectangles; (d) A sample of the decision tree for (c).

method that captured some interest for tackling MCDA problems, specially on the OR domain. DTs classify a pattern through a sequence of questions where the next question depends on the answer to the previous one. These trees are constructed as logical expressions as is illustrated in Fig. 4c-d). This ability generates a powerful data analysis tool capable to obtain interpretable results [38]. Nodes are consecutively split where a stop-splitting rule is required that controls the growth of the tree.

Neuron Networks are another kind of learning models. Multi-Layer Perceptron (MLP) is the most commonly used. A MLP is a layered structure consisting of nodes or units (called neurons) and one-way connections or links between the nodes of successive layers, such as the structure of Fig. 5a). The first layer is called the input layer, the last layer is the output layer, while the ones in the middle are called the hidden layers. Input layer of neurons is only a vector where all data are introduced triggering the learning process. Data propagates through the network in a forward direction, on a layer-by-layer basis. Layers are constituted by several neurons which commonly have non-linear and differentiable activation functions. Support Vector Machines (SVM) are another popular learning mechanism. In its simple form, SVMs uses a linear separating hyperplane to create a binary classifier

---

[1] The l_{0-1} *loss* function is the most commonly used one, i. e., $\mathcal{F}(a,b) = I(a \neq b)$ being I the identity function.

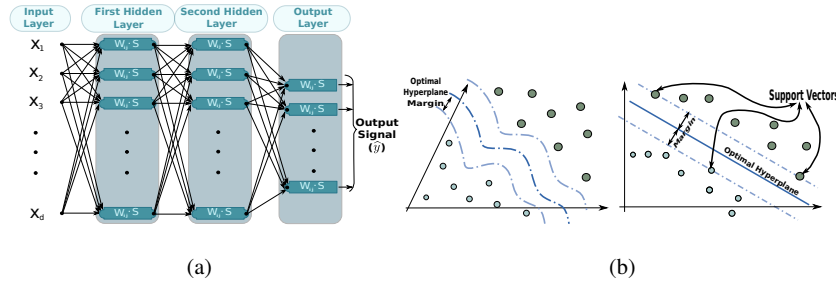(a)   (b)

Fig. 5: MLP and SVM methods: (a) Example of a MLP. This MLP is composed by 2 hidden layers, one input and output layer; (b) A two dimensional dataset is augmented to a higher feature space.

with a maximal margin. In cases where data cannot be linearly separable, data are transformed to a higher dimension than the original feature space—see Fig. 5b). Such is done by choosing a given kernel function, representing the inner product in some implicit higher dimension space. Formally, a kernel function is defined by $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}')$. This transformation ($\phi$) can be achieved by several non-linear mappings: e.g., polynomial, sigmoidal and radial basis functions. However, in a multiclass problem the usage of a binary SVM classifier can be limited. In order to improve this some heuristics and new formulations were proposed as an extension to the binary classification problem. Some of them encompass the OVO (One-Versus-One), OVA (One-Versus-All), DDAG (Decision Directed Acyclic Graph), single optimisation formulation, among others. Basically, OVO consists on the design of K(K-1)/2 binary classifiers where one class is discriminated against another. Similarly, and as the name suggests, OVA consists on the design of K binary classifiers where one class is compared against the others. Likewise the former heuristic, DAG, follow a similar procedure. The major difference is that prediction is made in a graph path manner where each node corresponds to a given binary classifier. In a completely different scenario, there are also techniques that try to define a single optimisation problem to solve the multiclass problem on SVMs.

This Section provided some key concepts regarding techniques for learning from data. Knowing that still much more has to be covered, the interested reader is advised to OR and AI textbooks [9, 38, 55, 65, 82, 108] for more information. Next Sections will describe different methods using some of the aforementioned methodologies for learning multicriteria models on ordinal data problems.

3 Multicriteria Decision Analysis

Multicriteria decision analysis (MCDA) is an important field within OR. It helped researchers to devise new approaches in order to analyse and interpret human's rea-

soning. Specifically, when handling several usually conflicting criteria towards an outcome. Such methods are generally composed by five phases depicted in Fig. 6.
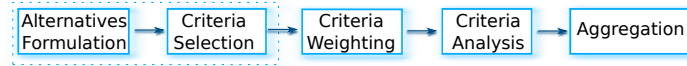

Fig. 6: Common Diagram of MCDA Methods [130, 137].

This Section will review multicriteria decision methods for ordinal data problems. Alternative formulation and criteria selection is usually defined by a set of experts or DMs [130, 137] and can depend on the problem in analysis. On the other hand, a given importance (weight) can be defined to each criterion whether through subjective or objective methods. After every criteria being considered, the analysis takes place. In general, MCDA tries to learn about users preferences encompassed in the different criteria considered. One key aspect of such methods is that they do not rely on any statistical assumptions [145]. Such highly contrasts with the topic which will be reviewed in Section 4. These two views can mark great differences on both topics, but as one shall see, there are points of connection between these two fields. In doing so, one can identify a trend towards filling the gap between OR and AI on MCDA. Finally, all criteria which were considered are then aggregated in order to define a plausible outcome.

It is important to stress that this work is mostly concerned with ordinal data. Hence, not all topics within MCDA can be covered in this Section. The interested reader is referred to [44, 58, 126, 147] for more information.

3.1 MCDA Methods

From Fig. 6, one can define methodologies which follow the whole process. Analytic Hierarchy Process (AHP) is one of such kind of frameworks [109]. After having the problem analysed and criteria selected, usually performed by an expert (or DM), it considers through an hierarchical approach each criteria [109]. However, recent reviews have argued that AHP results may not be the most desirable ones [62–64]. Mentioning that there is no clear evidence that AHP provides its users with their "best" decision [62], or in more detail, identifying the limitations in each step on the process [63]. Even though the Analytic Network Process (ANP) was introduced as a generalisation over AHP (a feedback network capable to adjust weights) [64, 110], few work has been done for the ordinal case.

ELECTRE [35, 107] and PROMETHEE [35, 37, 43] are two well known methods that, like AHP, can consist at most by the five steps illustrated in Fig. 6 [63]. Both techniques arose from the foundations of the outranking relation theory (ORT) [35]. In simple words, it consists of checking the outranking relation among instances

which permits to conclude whether an instance $\boldsymbol{x}^{(p)}$ outranks instance $\boldsymbol{x}^{(q)}$. Meaning, that instance $\boldsymbol{x}^{(p)}$ will be better for the DM than $\boldsymbol{x}^{(q)}$. This is achieved if there are enough statements to confirm (concordance) or to refute that (discordance). The two aforementioned methods require some preferential information which has to be defined by the DM. However, it may be difficult for the DM to understand the meaning of the preferences [61]. To overcome this, different improvements over the methods have been conducted. One of them was through the usage of evolutionary algorithms.

Evolutionary algorithms (EAs) came in a way to reproduce Darwin's theory of the survival of the fittest. EAs are also referred as populational meta-heuristics meaning that they work on the population space of solutions [13]. EAs generally encompasses on three major steps: 1) Gather a set of solutions; 2) Select a possible subset of candidates on that set of solutions and allow them to reproduce. Reproduction consists mainly on creating new solutions from the selected ones by crossover and mutation operators; 3) Finally, the process is repeated for the set of new solutions until a stopping criteria is achieved. Swiki in [116, 117] introduced an elitist evolutionary agent[2] system to solve multicriteria optimisation problems. By trying to reproduce biological mechanisms, an elitist group is introduced in the evolutionary architecture proposal. The final solution identified by the elitist group would indicate the desirable one which will dominate other possible solutions identified by other groups. Some hybrid approaches are also present in the literature [32, 42]. In [42] an outranking combined with an EA was proposed thanks to an indifference measure. Since preference modelling is cumbersome, authors used a population based meta-heuristic to generate the best solutions. An agent would then decide the best one. An approach proposed by Doumpos [32] comprehends the usage of concordance and discordance measures into a credibility index of an outranking method. This will assess the outranking relation among several alternatives. Since incomparable relations can occur, an EA is used to infer the parameters of the outranking method.

In a complete different setting, constraint programming tries to explore all possible combination of solutions thoroughly. Despite this being highly computational expensive, Junker in [66, 67] argues that an interactive approach has its advantages over state of the art techniques. It is also claimed that current existing methods do not express a clear explanation of the reason for one alternative being more preferable than another. In other words, a performance of 98% does not express which option is the best based on the original preferences. Using a special utility function to define preferences order in [67] a lexicographic optimal scheme is applied. Since lexicographic approach establish some ranking over the preferences order [41, 67], authors also permute the order of alternatives search. Bouveret [11] explores the idea in which characterises good solutions where multiple criteria have to be handled through the use of lexicographic algorithms.

Other methods incorporate cooperative algorithms which take part in the learning process from diverse sources of information and by different decision crite-

---

[2] In a simple way, an agent is a solution vector generated by some sub-optimal learning method.

ria [29, 71]. Methods with such properties are named Dominance-based Set Approach (DRSA) [29] which deal with the problem of multicriteria classification using maximum likelihood estimation. The problem is then solved by an optimal object reassignment algorithm. In [71] a stochastic DRSA approach is introduced. The rationale behind this method is to assess object class probability from an interval of classes.

Rough set theory is another field that one can count with when tackling MCDA. One interesting aspect is that rough set have the ability to produce a model of rule induction similar to data mining, knowledge discovery and machine learning [50]. In [50] authors extend the fuzzy set theory to rough sets theory in order to avoid as much as possible meaningless transformation of information. Rule induction is made through decision rules induced from dominance-based rough approximations of preference-ordered decision classes [51].

Let us now analyse in more depth contributions made to each node in the multicriteria methods process.

Criteria Weighting

Criteria weighting can be considered one of the most important steps for the decision maker. Once it weights the importance of each criterion, acting as a trade-off between criteria [61] that will be considered in the decision process, subtle changes can produce different outcome [136].

Methods for weighting criteria encompass equal weights, rank-order and hybrid approaches where after some considerations from the DM, weighting can be performed by a subjective or objective method [136, 137]. Equal weights ($w_j = 1/d$) is not valuable once relative importance among the criteria is ignored. Remains rank-order weighting approaches and their derivations to overcome these limitations. Another issue is that when dealing with uncertainty or incomplete information in any decision problem, the DM may not be reliable to define her/his preferences accurately. One way to handle this type of information is to represent preferences by a suitable distribution using stochastic multicriteria acceptability analysis (SMAA) methods. Several methods have been proposed in the literature— e.g. [40, 74, 75, 128] to name a few. SMAA-O proposed in [74] was an extension of SMAA works [127, 128] applied to ordinal (and cardinal) criteria. The problem is that, in the authors approach, an ordered criteria cannot be used directly in MC model. Therefore, it is assumed that exists a cardinal measure that corresponds to the known ordinal criteria and by considering consistent mappings between ordinal and cardinal scales, they randomly simulate such mapping through Monte Carlo iterations. Or in other words, ordinal data is converted into stochastic cardinal data by simulating consistent mappings between ordinal and cardinal scales that preserve the given labels. In SMAA literature review work of Tervonen [127] they claim that such simulations are not necessary since cardinal values can be interpreted directly.

Criteria Analysis

To the best of our knowledge, one of the first works in criteria analysis was proposed by Herstein [57] where an axiomatic approach was carried. A set of mathematical axioms was presented in this work to measure preferences order. Maccheroni in his work [85] explores the possibility where DM does not know for certain her/his preferences being therefore unable to rationalise her/his choices.

As previously mentioned, in the outranking approaches inconsistencies may arise when the preferences which are learned by given instances cannot be expressed through a model. Belacel in [6] proposes a construction of partial indifference indexes comparing pairs of preferences according to some criteria, aggregating them according to a concordance and non-discordance concept. Mousseau in [93] suggest to discard contradictory information from the preferences through an iterative aggregation-disaggregation scheme.

A number of variants of UTA [115] have been proposed in the literature over the last two decades and many works have been published concerned to this subject [8, 52, 54, 73, 146]. One related to ordinal problem was proposed in [145]. In this work, additive functions are used discriminating the preferences being evaluated from those that are not. Trying to go through a more natural way to human thinking over their outcomes or goals, some methods also based on utility functions have recently been proposed [88–90]. In this method, the authors developed a model to express logic of preferences in order to determine which of two outcomes is preferable.

Aggregation

As mentioned, aggregation models are one of the most studied methods within multicriteria decision analysis. For instance, in our credit scoring problem a model has to be designed to aggregate wage, payments history, age among others so that it can express the credit score profile of a given client. However, this approach implies that those functions have to be, among others, *monotone* [86]. Most important of all, the aggregation model has to be able to evince the importance of a criterion (done in the criteria analysis step), but also the interaction and compensation effects between criteria (done in the weighting step) [60]. Meaning that one has to design a model such that it can assign weights to a subset of possible criteria in order to capture these relations [60, 123].

As one saw until now, multicriteria methods encompass a variety of different approaches. Many of them address this problem through classification techniques using some sort of aggregation model [36, 44]. Afterwards, restrictions are then defined to the problem in question. However, despite the existence of the myriad of techniques, many pass through the definition of some objective function which can be delved through mathematical programming approaches. In [145] a multi-group hierarchical discrimination (M.H.DIS) method is defined. An error minimisation and clear group discrimination utility function is presented. Then, a two optimisa-

tion stages are conducted to avoid high computational complexity of MIP problems with many binary variables. An extension of this work is presented in [31] where the estimation of the additive utility functions in aforementioned work is accomplished through mathematical programming techniques. Two linear and one mixed-integer programs are used in M.H.DIS to estimate optimally the utility functions.

Unsupervised approaches such as the k-means algorithm or agglomerative hierarchical clustering (AHC) can also be used. The latter performs a hierarchical clustering where given individual clusters it can merge or split clusters until a stopping criteria is achieved. Given the utility matrix, authors employ clustering algorithms to form groups of alternatives (e.g., customers) with closely related preferences [77, 78]. However, in this phase little or no usage of the ordered criteria is explored.

4 Inductive Learning Algorithms

Inductive learning describes a very powerful field of research where machine learning (ML) lies. In ML one tries to obtain valid generalisation rules from data instead of the deductive learning approaches where one is already presented with a formalisation of the world and constructs (deducts) reasonable conclusions that cover our initial assumptions. Being also referred as a technique that *learns by example* (instances), it has been another thoroughly studied field which is composed by two main research topics: Regression and classification. A schematic of such problems and some real world scenarios are depicted in Fig. 7.
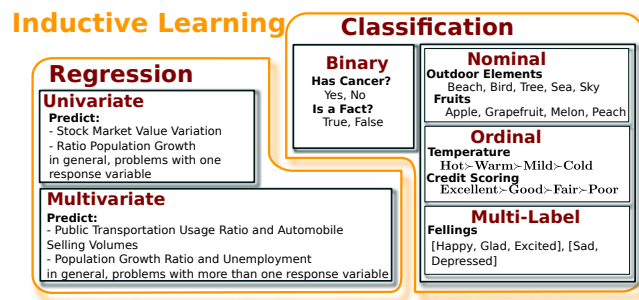

Fig. 7: Inductive Learning encompasses on two major research topics: *Regression* and *classification*. Both thrives on finding the best function that explains our data. The former renders the reasoning's on a continuous domain whereas the latter on a discrete (finite) domain. Each one is divided in other subtopics being their thoroughly analysis more appropriate for other textbooks [9, 38, 55] and here depicted just for context.

Learning mechanisms that solve ordinal problems have been tackled with both regression and classification strategies. Albeit being fundamentally different, both ordinal regression and ordinal classification methods have thrived among the scientific community, e.g., [18, 26, 46, 56, 72, 87, 113, 120], to name a few. The first works that tried to solve the classification of ordinal data were based on generalised linear models, as the cumulative model [87]. Tutz [129] presents a generic formulation for semi-parametric models extending the additive models [54]. In the machine learning community, Frank&Hall [46] have introduced a simple process which permits to explore information order in classification problems, using conventional binary classifiers as can be depicted in Fig. 8. In [56] it is applied the minimal structural risk principle [132] to derive a learning algorithm based in pairs of points.

Another way to learn ordering relation is by using classical algorithms of classification or regression and mapping the results into an ordinal scale. Kramer et al. [72] investigate the use of a learning algorithm for regression tasks—more specifically, a regression tree learner—to solve ordinal classification problems. In this case each class needs to be mapped to a numeric value. Kramer et al. [72] compare several different methods for doing this. However, if the class attribute represents a truly ordinal quantity—which, by definition, cannot be represented as a number in a meaningful way—there is no principled way of devising an appropriate mapping and this procedure is necessarily *ad hoc*. Harrington [53] argues that these type of approaches have many drawbacks as 1) makes regression learners sensitive to rank representation than their ordering and 2) since classification algorithms ignore rank order treating them as classes, it will be required more training data. Consequently, Harrington [53] presents a perceptron algorithm where its goal it to find a perceptron weight vector w which successfully projects all the instances into the k classes subintervals defined by some thresholds.

Moreover, existing methods incurring ordinal regression approaches fit data in general by a single rule defined by parts through K-1 thresholds [133]. This has a drawback since a mapping is required to convert ranks into real values or vice-versa. Hence, determining this mapping function is in general very difficult and makes regression learners more sensitive to rank value than their pairwise ordering. Some
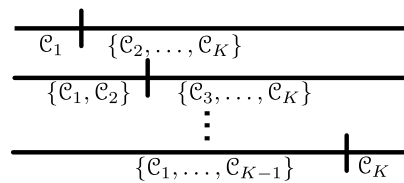

Fig. 8: Schematic of the proposal presented by Frank&Hall in [46]. Firstly it is performed a transformation of a K-class problem to a $K-1$ binary class problem. The training of the ith classifier involves the transformation of the K ordinal class into a binary one where the ith discriminator is obtained by separating the classes $C_1, \ldots, C_i$ and $C_{i+1}, \ldots, C_k$. The ith class represents the test $C_x > C_i$.

of the aforementioned drawbacks were avoided in Shashua and Levin [113] work where a generalised formulation of Support Vector Machines (SVMs) applied to ordinal data was proposed. However, such models can be too complex. Cardoso in [18] proposed a reduction technique to solve data ordinal problem classification using only one binary classifier. Following this idea, Lin et al. [83] explored the potential of solving ordinal problems through binary classification methods whereas Cheng et al. in [21] presented an adaptation of the Neural Networks (NN) towards ordinal problems. In [27] an order relation is incorporated among classes by imposing an unimodal distribution. This fundamental principle allowed to delve simpler Neural Networks (NNs) classifiers. The same rationale was instantiated to SVMs in [26] through the all-at-once strategy by solving a multiclass ordinal problem through a single optimisation process. Sun et al. in [124] proposed a Kernel Discriminant Analysis (KDA) for ordinal data. Even though authors argued that finding an optimal projection would result in better reasonable results, in doing so one would loose its relation to the original features. Hence, in the case of need for interpretable results, through the usage of such methods, one would be unable to understand the reason of the outcome given specific features.

Metric learning is research subject that recently has been gaining increasingly attention, specially in the machine learning community [138, 142, 144]. The performance of all machine learning algorithms depends critically on the metric that is used over the input space. Some learning algorithms, such as K-means and k-nearest neighbours, require a metric that will reflect important relationships between each classes in data and will allow to discriminate instances belonging to one class from others [104]. Ouyang [96, 111] explored this subject in the ordinal problem. In [96] by assuming that closer instances in the input space should translate an order of relation, a metric distance is learn so that pairs of instances are closer than the remainder pairs. However, class label is discarded in this approach.

Other approaches [22–24, 143] consisted on probabilistic approaches based in Gaussian processes to learn models for the ordinal problem. In [143] a collaborative approach is delved towards better, not only in accuracy, but also in a context of collaborative preference learning.

Regarding decision trees (DTs) for ordinal data, some works consider problems that are monotone, i.e., all attributes have ordered domains. Meaning, if $\boldsymbol{x}, \boldsymbol{z}$ are data points such that $\boldsymbol{x} \leq \boldsymbol{z}$ ($x_i \leq z_i$ for each criteria i) then their classes should satisfy the condition $\widehat{f}(\boldsymbol{x}) \leq \widehat{f}(\boldsymbol{z})$, where $\widehat{f}(.)$ is the labeling function. Potharst [99–101] proposes a method that induces a binary decision tree from a monotone dataset. Other methods were also proposed for non-monotone datasets (the most likely scenario in the presence of noise) where the resulting tree may be non-monotone. In this scenario, a fuzzy operator was used instead of a entropy function for performance measurement [30]. Works on k-nearest neighbour for ordinal data seems even scarcer. Besides the well-known adaptation of using the median as labelling rule instead of mode for the k labels, literature only presents a modified version of the standard k-NN for the construction of monotone classifiers from data [39]. Again, this work continues to be limited by the assumption of monotonocity in the input data. In general, the monotone constraint was overcame in [19, 120]. Argu-

ing that ordinality could not be captured directly from the input space, but from the feature space, authors explored a re-labelling approach on the output decision space through a postprocessing optimisation procedure.

From the works until now revised, one has encountered several methods that make use of different procedures from operations research field, and other proposals design their learning models so that multicriteria can be rendered in the learning phase. In this setting, multicriteria assessment is simply performed over a set of diverse unattached reasoning's which renders the desirable outcomes without a clear understanding of which criteria contributed most. To overcome this, De Smet et al. [118] developed a k-means clustering algorithm in a multicriteria decision analysis perspective.

In this section we have reviewed several learning approaches for the resolution of the ordinal problem. In the end, it is obvious how increasingly this subject has been studied. The reasons can be due to the awareness of its transversal usability in a set of diverse applications. However, due to the background of many researchers, many have tried to solve this problem through regression, classification and ranking methodologies. The work of Furnkranz et al. [48, 49] despite using a pairwise approach, compared ranking and classification principles in their proposals. As final remark, one must note how vastly such methods can be employed such it has been explored by Van Belle et al. [114, 131]. In these works, different approaches have been delved towards ranking, ordinal and survival analysis problems. Even though authors performed strict assumptions on data to develop their models, such as monotone data, it still is a good example of the importance of this topic in the inductive learning field.

4.1 Feature Selection Algorithms on Ordinal Data

Nowadays, it is relatively easy to solve problems with millions of instances, each of them with a reasonable number of features. However, it is common to have access to datasets with significantly higher number of features than instances leading to the well known problem of the curse of dimensionality. Feature selection (FS) techniques provide the means to overcome this issue by identifying the most valuable features so that good and simple class discrimination models can be obtained. Furthermore, a noise reduced dataset can be achieved since these methods can "clean" data from features with noise [34].

There are three types of feature selection algorithms: Filter, wrapper and embedded. The former is independent of the classifier being usually done before the learning phase. Wrapper algorithms iteratively select subset of features and assess the learning models performance to determine how useful that set of features are whereas embedded algorithms select automatically features during the model construction [34, 106]. Fig. 9 succinctly depicts the three approaches.

Feature selection on ordinal data is a relatively young topic. In [84] a χ^2 statistic method is used to discretize numeric features as a way to select features. Even
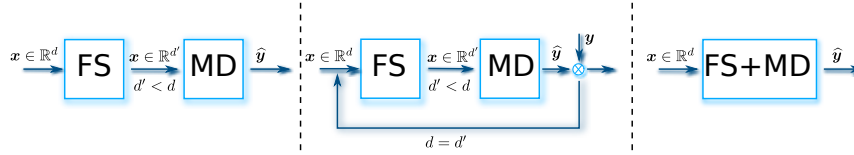
Fig. 9: Three different standard approaches for feature selection: (left) depicts the *filter* feature selection (FS) approach done before the model design (MD); (centre) the *wrapper* is consisted on an iterative approach where features are removed step by step until a desirable performance of the model is achieved; and (right) *embedded* method is designed jointly with the learning model algorithm.

though the method proposed by Liu [84] was identified as being limited to a first-order feature-class correlation (i.e., are linearly correlated), such should not be seen as a drawback. Once highly complex learning machines could easily cope with the data complexity and infer a linear relation with the features and classes, or more precisely, perform overfitting on data [112, 121]. Nevertheless, Last et al in [79] proposed an information-theoretic method for feature selection by performing a discretization over the features in order to minimise classes entropy. Even though ordinal data can contain only discrete features fitting well to this technique, there are datasets with continuous features (see for instance [17]). In such scenarios, applying a discretization technique can lead to loss of accuracy in the model design. Despite being mentioned the capability to handle ordinal data, no experiment has been conducted, neither their methods were designed for this type of problems. Through a completely different approach, Xia et al. [140] presents a recursive approach to extract features where it learns consecutively new rules from instances represented by the new features.

Other techniques in the ordinal context have been referred to Baccianella et al in [4, 5]. Using only the filter approach for feature selection, authors used several measures to identify feature relevance through the minimisation of the instances variance over all classes, similarity, information gain and negative correlation according to the class label, specifically developed for ordinal problems. Finally, Sousa et al. [121] explored a concept introduced by Rodriguez et al. [106] where they tackle the FS problem in one-step process through quadratic programming as represented in Equation (6). The quadratic term (Q in Equation (6)) would capture the redundancy whereas the linear term (F in Equation (6)) would capture the relevance.

$$\min_{\boldsymbol{x}} \left\{ \frac{1}{2}(1-\alpha)\boldsymbol{x}^t Q\boldsymbol{x} - \alpha F^t \boldsymbol{x} \right\} \tag{6}$$

Here α is the trade-off between relevance and redundancy which can be empirically defined. In order to capture the ordinal relation on data in this setting, authors chosen the Minimum Spanning Trees (MST) as the linear term (F) to assess the increase of complexity when a subset of features is removed. However, one of the issues

identified in this approach concerns to the fact that authors did not take advantage of the ordinal information that could be explicitly included on data (quadratic term).

4.2 Performance Measures

After considering the advantages and disadvantages, goals achieved and open issues of the techniques presented in previous sections, the discussion of how to measure the performance of such techniques has not been debated much.

Usually, a learning process consists in two main phases: A cross-validation phase and an estimation of the model performance ($\mathcal{F}$ represented in Equation (5)) on a real-world scenario (also known as the testing phase). In both situations, one has to analyse the performance of a model given certain parametrization and its behaviour in a non controllable environment, respectively. Herein, the question that one obviously poses is: How much did the model err? Or, how much the prediction differs from the real outcome? Given certain assumptions of models design, it is clear, as we will shortly show, that the metric chosen for this task is crucial.

It is interesting to see that in contrast to the plethora of existing methods concerning multicriteria learning, only recently we witnessed some concerns to this issue [20, 47, 81], disregarding advances performance made on the broader field of machine learning [80]. Knowing that "no free lunch" theorems state that there is not an algorithm that can be superior on all problems in regard to classification accuracy [139], the assessment of an appropriate learning method given a specific problem is desirable [80].

For classification problems, MER (Misclassification Error Rate) is currently one of the most used measures. Its widely use make it a *de facto* standard when comparing different learning algorithms by just counting the misclassifications occurred. In other problems domains, it is usual to penalise the misclassifications by weighting them by the magnitude of the error to avoid uneven results. When such happens, MAE (Mean Absolute Error) and MSE (Mean Squared Error) measures are usually the most appropriate choices. Summing, the performance of a classifier can be assessed in a dataset $\mathcal{D}$ through

$$\frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{D}}|g(\mathcal{C}_{\boldsymbol{x}})-g(\widehat{\mathcal{C}}_{\boldsymbol{x}})| \qquad \frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{D}}\left(g(\mathcal{C}_{\boldsymbol{x}})-g(\widehat{\mathcal{C}}_{\boldsymbol{x}})\right)^2,$$

respectively, where $g(.)$ corresponds to the number assigned to a class, $N = \text{card}(\mathcal{D})$, and $\mathcal{C}_{\boldsymbol{x}}$ and $\widehat{\mathcal{C}}_{\boldsymbol{x}}$ are the true and estimated classes. However, this assignment is arbitrary and the numbers chosen to represent the existing classes will evidently influence the performance measurement given by MAE or MSE. A clear improvement on these measures would be to define them directly from the confusion matrix CM (a table with the true class in rows and the predicted class in columns, with each entry $n_{r,c}$ representing the number of instances from the $r-$th class predicted as being from $c-$th class):

$$MAE = \frac{1}{N} \sum_{r=1}^{K} \sum_{c=1}^{K} n_{r,c} |r - c| \qquad MSE = \frac{1}{N} \sum_{r=1}^{K} \sum_{c=1}^{K} n_{r,c} (r - c)^2$$

where K is the number of classes. We will always assume that the ordering of the columns and rows of the CM is the same as the ordering of the classes. This procedure makes MAE and MSE independent of the numbers or labels chosen to represent the classes. To a certain degree, these two measures are better than MER because they take values which increase with the absolute differences between 'true' and 'predicted' class numbers and so the misclassifications are not taken as equally costly.

In order to avoid the influence of the numbers chosen to represent the classes on the performance assessment, it has been argued that one should only look at the order relation between 'true' and 'predicted' class numbers. The use of Spearman's rank correlation coefficient, R_s, and specially Kendall's tau-b, τ_b, is a step in that direction [70, 122]. For instance, in order to compute R_s, we start by defining two rank vectors of length N which are associated with the variables $g(\mathcal{C})$ and $g(\widehat{\mathcal{C}})$. There will be many examples in the dataset with common values for those variables; for these cases average ranks are used. If $\boldsymbol{p}$ and $\boldsymbol{q}$ represent the two rank vectors, then $R_s = \frac{\Sigma(p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\Sigma(p_i - \bar{p})^2 \Sigma(q_i - \bar{q})^2}}$. As we can see, Spearman's coefficient is still dependent on the values chosen for the ranks representing the classes and so it is not completely appropriate to measure the performance of ordinal data classifiers. More importantly, R_s looses information about the absolute value of the classes. Kendall's coefficient τ_b has been advocated as a better measure for ordinal variables because it is independent of the values used to represent classes [70]. Its robustness is achieved by working directly on the set of pairs corresponding to different observations. However, there are limitations: By working only with the relative order of elements, it loses information about the absolute prediction for a given observation.

Other attempts have considered the analysis of the learner behaviour on a ROC (Receiver Operating Characteristic) curve or its equivalent, AUC (Area Under Curve). Despite empirical evidences of AUC providing more desirable properties when compared to accuracy [12] only recently this topic was not only re-proposed but also new evidences of its advantages were shown [59]. In this work, AUC is demonstrated as an objective measure for selecting the best learning model, but, and most important, refers to the need of developing better measures for learner design and performance assessment [59]. In this line of research, in [134] it is compared different ROC measurements. However, and despite the assumptions made, ROC derived measures that assess a ranking for different performance do not quantify the performance achieved by a learner [133]. Such analysis, although with different purposes, has been conducted by [7] using Cohen's kappa statistic.

On the other way, the discussion was revamped by Baccianella et al [3] through an analysis of different derivations of MSE and MAE metrics for ordinal problems. This work is key since it debates two main issues incurred on the performance measurement of learners for this type of classification problems: Imbalanced classes and classes with equal penalisation costs. In order to avoid the former problematic,

a derivation from MAE is presented by averaging the deviations per class.

$$MAE^M = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{g(\widehat{\mathcal{C}_i})} |g(\mathcal{C}_i) - g(\widehat{\mathcal{C}_i})|$$

In the same line, the coefficient r_{int} was recently introduced, taking into account the expected high number of ties in the values to be compared [27]. In fact, the variables $\mathcal{C}$ and $\widehat{\mathcal{C}}$ are two special ordinal variables. Because there are usually very few classes compared to the number of observations, these variables will take many tied values (most of them, in fact). Nevertheless, r_{int} is sufficiently general and, if there are no tied values, it can still be applied as it is. Like τ_b, r_{int} assumes that the only thing that matters is the order relation between such values, which is the same as the order relation between the classes. This coefficient takes values in $[-1, 1]$, in contrary to MAE (and MSE) which are upper-unbounded. The latter can be identified as a limitation. Another observation is that it is fair to compare MAE results in two different applications with a different number of observations, N, since MAE is properly normalised by N. However, if the applications involve a different number of classes, K, it is not clear how to compare the performance obtained in the two settings.

In [20] a different approach was taken. Even though the adaptation of the MAE and MER to a confusion matrix form surpasses standard forms, there are still issues regarding these metrics. Some of the vicissitudes as mentioned in [20] encompass: Equally misclassification costs, metrics unable to evaluate example dependent costs [14] or metrics more proper to ranking problems. Having Cardoso and Sousa identified some key advantages of using the CM form, and given the merit of both MAE and MER fundamental principles, they proposed a new one that takes advantage of all as a single optimisation problem. This new metric chooses pairs of observations from the CM that do not contradict the relative order given by the classifier and the true relative class order which minimise the cost of a global optimisation procedure. The choice is done in a way that minimises the deviation of the pairs to the main diagonal while maximising the entries values in the path that cross the CM. This was formalised as

$$OC_\beta^\gamma = \min \left\{ 1 - \frac{\sum_{(r,c) \in \text{path}} n_{r,c}}{N + \left(\sum_{\forall(r,c)} n_{r,c} |r - c|^\gamma \right)^{1/\gamma}} + \beta \sum_{(r,c) \in \text{path}} n_{r,c} |r - c|^\gamma \right\}, \quad (7)$$

where the minimisation is performed over all possible paths from (1,1) to (K,K). γ and β were defined based upon empirical experiments.

Other techniques can also go through data generators methodologies where one can control the statistical properties herein aiding in the learners benchmark [47]. More importantly, techniques capable to manipulate Bayes error rate can foster new lines of research where fair learners comparison [7] and the development of new ones can take place.

In [20] it is raised a question that interesting enough has not been debated since [81] in the ordinal context. As one knows, the usage of such metrics in the

design of classifiers can be done on two distinct situations. A first use is 'externally' to the classifier, using the metric to select the best parametrization of the classifier (usually when performing a cross-validation procedure). A second possibility is to embed the new metric in the classifier design, adapting the internal objective function of the classifier, replacing loss functions based on standard measures by a loss function based on the proposed measure. For instance, the standard loss function of a neural network based on the square of the error or on cross-entropy could be replaced by an error evaluated by an appropriate metric [59]. Lee [81] accomplished such for the design of ordinal trees, but since then few works have addressed this subject in the ordinal context.

It is interesting that only recently we saw a significant growth of the awareness of this topic importance. Even though some works have already tackled this issue, all lack on concretely assessing the performance of a given ordinal learning model. Until now, new metrics have been designed and compared against MAE followed by some reasoning. The problem resides how close a metric is in expressing accuracy. Different prosaically strategies can pass through the definition of prior costs for each class [95] or, when using a given set of different metrics, a meta-metric to assess the performance of metrics should be in place as suggested by Cardoso [20].

5 Conclusion

Multicriteria (MC) has been studied for over more than five decades where recent years presented interesting developments. Aside novel methodologies, a trend towards the generalisation of this problem was identified where at the same time a new light was shed over this topic thanks to a niche of applications. In this chapter a thoroughly review was conducted on two major disciplines: Operations research (OR) and artificial intelligence (AI).

MCDA has a strong connection with OR community. Fuzzy Set theory research community was one that rapidly proposed new models towards these problems. Their capability to handle uncertainty can be identified as an asset in these models. Even though in other research fields MC is giving its first steps, a new trend is appearing as a number of different studies are taking place. On the other hand, evolutionary approaches are still on the very beginning regarding ordinal problems. It also has been claimed that some approaches do not cope well with many criteria or do not capture correctly every rationale taken by the decision maker.

In the AI domain, it was described that albeit the myriad of techniques, some do not totally incorporate or effectively use the additional information of order in the classifier construction. Others have a higher complexity to be useful in real problems or require specific optimisation algorithms during the classifier construction. Also, it was identified that is still common the usage of regression approaches to solve the ordinal data problem. Notwithstanding, some improvements have been achieved. Simplifications have been introduced through the usage of a standard binary classification techniques and fundamental principles towards the ordinal data

problem. Such theories have proved to be valuable in the design of simpler classi-fiers and when not possible, in the design of posterior rules to impose ordinality. Another question that has recently been tackled concerns about finding good met-rics for measuring learners performance. We reviewed many adaptations of standard metrics and new ones that optimise different criteria of the learner behaviour.

In the end, and in spite of much of what has been achieved, a fair comparison between methods of both fields is still lacking. It was also clear that MC is very rich in terms of nomenclature. Having identified what has been achieved and current open issues, it is expected that this study leads to future technical developments and topic convergence.

References

1. Abraham, A., Abraham, A., Falcn, R., Bello, R.: Rough Set Theory: A True Landmark in Data Analysis. Springer Publishing Company, Incorporated (2009)
2. Angilella, S., Greco, S., Matarazzo, B.: Non-additive robust ordinal regression: A multiple criteria decision model based on the choquet integral. European Journal of Operational Research **201**(1), 277 – 288 (2010). DOI 10.1016/j.ejor.2009.02.023
3. Baccianella, S., Esuli, A., Sebastiani, F.: Evaluation measures for ordinal regression. In: Proceedings of the Ninth International Conference on Intelligent Systems Design and Appli-cations, pp. 283–287 (2009)
4. Baccianella, S., Esuli, A., Sebastiani, F.: Feature selection for ordinal regression. In: Pro-ceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, pp. 1748–1754. ACM, New York, NY, USA (2010). DOI 10.1145/1774088.1774461
5. Baccianella, S., Esuli, A., Sebastiani, F.: Selecting features for ordinal text classification. In: IIR, pp. 13–14 (2010)
6. Belacel, N.: Multicriteria assignment method PROAFTN: Methodology and medical ap-plication. European Journal of Operational Research **125**(1), 175–183 (2000). DOI 10.1016/S0377-2217(99)00192-7
7. Ben-David, A.: A lot of randomness is hiding in accuracy. Engineering Applications of Artificial Intelligence **20**(7), 875–885 (2007). DOI 10.1016/j.engappai.2007.01.001
8. Beuthe, M., Scannella, G.: Comparative analysis of UTA multicriteria methods. European Journal of Operational Research **130**(2), 246–262 (2001). DOI 10.1016/S0377-2217(00)00042-4
9. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statis-tics), 1st ed. 2006. corr. 2nd printing edn. Springer (2007)
10. Blaszczynski, J., Greco, S., Slowinski, R., Szelg, M.: Monotonic variable consistency rough set approaches. International Journal of Approximate Reasoning **50**(7), 979–999 (2009). DOI 10.1016/j.ijar.2009.02.011. Special Section on Graphical Models and Information Re-trieval
11. Bouveret, S., Lemaître, M.: Computing leximin-optimal solutions in constraint networks. Artificial Intelligence **173**(2), 343–364 (2009). DOI 10.1016/j.artint.2008.10.010

12. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145 – 1159 (1997). DOI 10.1016/S0031-3203(96)00142-2
13. Branke, J., Deb, K., Miettinen, K., Slowinski, R. (eds.): Multiobjective Optimization: Interactive and Evolutionary Approaches. Springer-Verlag, Berlin, Heidelberg (2008)
14. Brefeld, U., Geibel, P., Wysotzki, F.: Support vector machines with example dependent costs. In: N. Lavrac, D. Gamberger, H. Blockeel, L. Todorovski (eds.) Machine Learning: ECML 2003, *Lecture Notes in Computer Science*, vol. 2837, pp. 23–34. Springer Berlin / Heidelberg (2003)
15. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: Classification and Regression Trees. Chapman & Hall (1998)
16. Cao-Van, K., De Baets, B.: Consistent representation of rankings. In: H. de Swart, E. Orlowska, G. Schmidt, M. Roubens (eds.) Theory and Applications of Relational Structures as Knowledge Instruments, *Lecture Notes in Computer Science*, vol. 2929, pp. 1966–1967. Springer Berlin / Heidelberg (2003). DOI 10.1007/978-3-540-24615-2_6
17. Cardoso, J.S., Cardoso, M.J.: Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artificial Intelligence in Medicine **40**, 115–126 (2007)
18. Cardoso, J.S., Costa, J.F.P.d.: Learning to classify ordinal data: the data replication method. Journal of Machine Learning Research **8**, 1393–1429 (2007)
19. Cardoso, J.S., Sousa, R.: Classification models with global constraints for ordinal data. In: Proceedings of The Ninth International Conference on Machine Learning and Applications (ICMLA) (2010)
20. Cardoso, J.S., Sousa, R.: Measuring the Performance of Ordinal Classification. International Journal of Pattern Recognition and Artificial Intelligence (2011)
21. Cheng, J., Wang, Z., Pollastri, G.: A neural network approach to ordinal regression. In: Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pp. 1279 –1284 (2008). DOI 10.1109/IJCNN.2008.4633963
22. Chu, W., Ghahramani, Z.: Gaussian Processes for Ordinal Regression. J. Mach. Learn. Res. **6**, 1019–1041 (2005)
23. Chu, W., Ghahramani, Z.: Preference learning with Gaussian processes. In: ICML '05: Proceedings of the 22nd international conference on Machine learning, pp. 137–144. ACM, New York, NY, USA (2005). DOI 10.1145/1102351.1102369
24. Chu, W., Sindhwani, V., Ghahramani, Z., Keerthi, S.S.: Relational Learning with Gaussian Processes. In: B. Schölkopf, J. Platt, T. Hoffman (eds.) Advances in Neural Information Processing Systems 19, pp. 289–296. MIT Press, Cambridge, MA (2007)
25. Cossock, D., Zhang, T.: Subset ranking using regression. In: G. Lugosi, H. Simon (eds.) Learning Theory, *Lecture Notes in Computer Science*, vol. 4005, pp. 605–619. Springer Berlin / Heidelberg (2006). DOI 10.1007/11776420_44
26. da Costa, J.F.P., Sousa, R., Cardoso, J.S.: An all-at-once unimodal svm approach for ordinal classification. In: Proceedings of The Ninth International Conference on Machine Learning and Applications (ICMLA) (2010)
27. Costa, J.F.P.d., Alonso, H., Cardoso, J.S.: The unimodal model for the classification of ordinal data. Neural Networks **21**(1), 78–91 (2008)
28. Delannay, N., Verleysen, M.: Collaborative filtering with interlaced generalized linear models. Neurocomputing **71**(7-9), 1300–1310 (2008). DOI http://dx.doi.org/10.1016/j.neucom.2007.12.021
29. Dembczynski, K., Greco, S., Kotlowski, W., Slowinski, R.: Statistical Model for Rough Set Approach to Multicriteria Classification. In: PKDD 2007: Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases, pp. 164–175. Springer-Verlag, Berlin, Heidelberg (2007)
30. Dombi, J., Zsiros, A.: Learning multicriteria classification models from examples Decision rules in continuous space. European Journal of Operational Research **160**(3), 663–675 (2005). DOI 10.1016/j.ejor.2003.10.006. Decision Analysis and Artificial Intelligence

31. Doumpos, M., Kosmidou, K., Baourakis, G., Zopounidis, C.: Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis. European Journal of Operational Research **138**(2), 392–412 (2002). DOI 10.1016/S0377-2217(01)00254-5

32. Doumpos, M., Marinakis, Y., Marinaki, M., Zopounidis, C.: An evolutionary approach to construction of outranking models for multicriteria classification: The case of the electre tri method. European Journal of Operational Research **199**(2), 496–505 (2009). DOI 10.1016/j.ejor.2008.11.035

33. Doumpos, M., Pasiouras, F.: Developing and testing models for replicating credit ratings: A multicriteria approach. Computational Economics **25**, 327–341 (2005)

34. Doumpos, M., Salappa, A.: Feature selection algorithms in classification problems: an experimental evaluation. In: Proceedings of the 4th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering Data Bases, pp. 36:1–36:6. World Scientific and Engineering Academy and Society (WSEAS) (2005)

35. Doumpos, M., Zopounidis, C.: Multicriteria Decision Aid Classification Methods. Kluwer Academic Publishers, Dordrecht (2002)

36. Doumpos, M., Zopounidis, C.: A multicriteria classification approach based on pairwise comparisons. European Journal of Operational Research **158**(2), 378–389 (2004). DOI 10.1016/j.ejor.2003.06.011. Methodological Foundations of Multi-Criteria Decision Making

37. Doumpos, M., Zopounidis, C.: A multicriteria decision support system for bank rating. Decision Support Systems **50**(1), 55 – 63 (2010). DOI 10.1016/j.dss.2010.07.002

38. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Edition), 2 edn. Wiley-Interscience (2001)

39. Duivesteijn, W., Feelders, A.: Nearest neighbour classification with monotonicity constraints. In: ECML PKDD '08: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I, pp. 301–316. Springer-Verlag, Berlin, Heidelberg (2008)

40. Durbach, I.N.: The use of the SMAA acceptability index in descriptive decision analysis. European Journal of Operational Research **196**(3), 1229–1237 (2009). DOI 10.1016/j.ejor.2008.05.021

41. Ehrgott, M.: Multicriteria optimization. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag (2000)

42. Fernandez, E., Navarro, J., Bernal, S.: Multicriteria sorting using a valued indifference relation under a preference disaggregation paradigm. European Journal of Operational Research **198**(2), 602–609 (2009). DOI 10.1016/j.ejor.2008.09.020

43. Figueira, J., Greco, S., Ehrogott, M., Brans, J.P., Mareschal, B.: Promethee methods. In: Multiple Criteria Decision Analysis: State of the Art Surveys, *International Series in Operations Research & Management Science*, vol. 78, pp. 163–186. Springer New York (2005). DOI 10.1007/0-387-23081-5_5

44. Figueira, J., Greco, S., Ehrogott, M., Siskos, Y., Grigoroudis, E., Matsatsinis, N.: Uta methods. In: F.S. Hillier (ed.) Multiple Criteria Decision Analysis: State of the Art Surveys, *International Series in Operations Research & Management Science*, vol. 78, pp. 297–334. Springer New York (2005). DOI 10.1007/0-387-23081-5_8

45. Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics **7**, 179–188 (1936)

46. Frank, E., Hall, M.: A Simple Approach to Ordinal Classification. In: EMCL' 01: Proceedings of the 12th European Conference on Machine Learning, pp. 145–156. Springer-Verlag, London, UK (2001)

47. Frasch, J.V., Lodwich, A., Shafait, F., Breuel, T.M.: A Bayes-true data generator for evaluation of supervised and unsupervised learning methods. Pattern Recognition Letters **32**(11), 1523–1531 (2011). DOI 10.1016/j.patrec.2011.04.010

48. Fürnkranz, J., Hüllermeier, E.: Pairwise Preference Learning and Ranking. In: Proceedings of the 14th European Conference on Machine Learning, pp. 145–156. Springer-Verlag (2003)

49. Fürnkranz, J., Hüllermeier, E.: Pairwise preference learning and ranking. Tech. rep., Austrian Research Institute for Artificial Intelligence, Wien, Austria (2003)

50. Greco, S., Inuiguchi, M., Slowinski, R.: Fuzzy rough sets and multiple-premise gradual decision rules. International Journal of Approximate Reasoning **41**(2), 179–211 (2006). DOI 10.1016/j.ijar.2005.06.014. Advances in Fuzzy Sets and Rough Sets

51. Greco, S., Matarazzo, B., Slowinski, R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research **129**(1), 1–47 (2001). DOI 10.1016/S0377-2217(00)00167-3

52. Greco, S., Mousseau, V., Slowinski, R.: Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. European Journal of Operational Research **191**(2), 416–436 (2008). DOI 10.1016/j.ejor.2007.08.013

53. Harrington, E.F.: Online Ranking/Collaborative Filtering Using the Perceptron Algorithm. In: Proceedings of the 20th International Conference on Machine Learning, pp. 250–257 (2003)

54. Hastie, T., Tibshirani, R.: Generalized Additive Models. Statistical Science **1**, 297–318 (1986)

55. Haykin, S.: Neural Networks and Learning Machines (3rd Edition), 3 edn. Prentice Hall (2008)

56. Herbrich, R., Graepel, T., Obermayer, K.: Regression Models for Ordinal Data: A Machine Learning Approach. Tech. rep. (1999)

57. Herstein, I.N., Milnor, J.: An axiomatic approach to measurable utility. Econometrica **21**(2), 291–297 (1953)

58. Hillier, F.S., Lieberman, G.J., Hillier, F., Lieberman, G.: MP Introduction to Operations Research. McGraw-Hill Science/Engineering/Math (2004)

59. Huang, J., Ling, C.: Using auc and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on **17**(3), 299 – 310 (2005). DOI 10.1109/TKDE.2005.50

60. Huédé, F., Grabisch, M., Labreuche, C., Savéant, P.: Integration and propagation of a multicriteria decision making model in constraint programming. Journal of Heuristics **12**(4-5), 329–346 (2006). DOI 10.1007/s10732-006-8075-2

61. Iryna, Y.: Solving classification problems with multicriteria decision aiding approaches (2007)

62. Ishizaka, A., Balkenborg, D., Kaplan, T.: Does ahp help us make a choice? an experimental evaluation. JORS **62**(10), 1801–1812 (2011). DOI 10.1057/jors.2010.158

63. Ishizaka, A., Labib, A.: Analytic hierarchy process and expert choice: Benefits and limitations. OR Insight **22**(4), 201–220 (2009). DOI 10.1057/ori.2009.10

64. Ishizaka, A., Labib, A.: Review of the main developments in the analytic hierarchy process. Expert Systems with Applications **38**(11), 14,336 – 14,345 (2011). DOI 10.1016/j.eswa.2011.04.143

65. Jensen, R., Shen, Q.: Computational intelligence and feature selection: Rough and fuzzy approaches (2008)

66. Junker, U.: Preference-based search and multi-criteria optimization. Annals of Operations Research **130**(1), 75—115 (2004)

67. Junker, U.: Preference-based problem solving for constraint programming pp. 109–126 (2008). DOI 10.1007/978-3-540-89812-2_8

68. Kangas, J., Kurttila, M., Kajanus, M., Kangas, A.: Evaluating the management strategies of a forestland estate–the s-o-s approach. J. Environ. Manage. **69**(4), 349–58 (2003). DOI 10.1016/j.jenvman.2003.09.010

69. Kecman, V.: Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. MIT Press, Cambridge, MA, USA (2001)

70. Kendall, M.: A new measure of rank correlation. Biometrika **30**, 81–89 (1938)

71. Kotlowski, W., Dembczynski, K., Greco, S., Slowinski, R.: Stochastic dominance-based rough set model for ordinal classification. Information Sciences **178**(21), 4019–4037 (2008)

72. Kramer, S., Widmer, G., Pfahringer, B., De Groeve, M.: Prediction of Ordinal Classes Using Regression Trees. Fundam. Inf. **47**(1-2), 1–13 (2001)

73. Köksalan, M., Özpeynirci, S.B.: An interactive sorting method for additive utility functions. Computers & Operations Research **36**(9), 2565–2572 (2009). DOI 10.1016/j.cor.2008.11. 006

74. Lahdelma, R., Miettinen, K., Salminen, P.: Ordinal criteria in stochastic multicriteria acceptability analysis (smaa). European Journal of Operational Research **147**(1), 117–127 (2003). DOI 10.1016/S0377-2217(02)00267-9

75. Lahdelma, R., Salminen, P.: Prospect theory and stochastic multicriteria acceptability analysis (SMAA). Omega **37**(5), 961–971 (2009). DOI 10.1016/j.omega.2008.09.001

76. Lahdelma, R., Salminen, P., Hokkanen, J.: Locating a waste treatment facility by using stochastic multicriteria acceptability analysis with ordinal criteria. European Journal of Operational Research **142**(2), 345 – 356 (2002). DOI 10.1016/S0377-2217(01)00303-4

77. Lakiotaki, K., Delias, P., Sakkalis, V., Matsatsinis, N.: User profiling based on multi-criteria analysis: the role of utility functions. Operational Research **9**, 3–16 (2009). DOI 10.1007/s12351-008-0024-4

78. Lakiotaki, K., Matsatsinis, N., Tsoukià ands, A.: Multicriteria user modeling in recommender systems. Intelligent Systems, IEEE **26**(2), 64 –76 (2011). DOI 10.1109/MIS.2011.33

79. Last, M., Kandel, A., Maimon, O.: Information-theoretic algorithm for feature selection. Pattern Recognition Letters **22**(6-7), 799 – 811 (2001). DOI 10.1016/S0167-8655(01)00019-8

80. Lavesson, N., Davidsson, P.: Evaluating learning algorithms and classifiers. Int. J. Intell. Inf. Database Syst. **1**, 37–52 (2007). DOI 10.1504/IJIIDS.2007.013284

81. Lee, J., Liu, D.Z.: Induction of ordinal decision trees. In: Proceedings of the International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2220–2224 (2002)

82. Lee, K.H.: First Course On Fuzzy Theory And Applications. SpringerVerlag (2004)

83. Lin, H.T., Li, L.: Combining ordinal preferences by boosting. In: Proceedings ECML/PKDD 2009 Workshop on Preference Learning, pp. 69–83 (2009)

84. Liu, H., Setiono, R.: Feature selection via discretization. Knowledge and Data Engineering, IEEE Transactions on **9**(4), 642 –645 (1997). DOI 10.1109/69.617056

85. Maccheroni, F., Marinacci, M., Rustichini, A.: Ambiguity aversion, robustness, and the variational representation of preferences. Econometrica **74**(6), 1447–1498 (2006)

86. Marichal, J.L.: Aggregation Operators for Multicriteria Decision Aid. Ph.D. thesis, Institute of Mathematics, University of Liège, Liège, Belgium (1998)

87. McCullagh, P.: Regression Models for Ordinal Data. Journal of the Royal Statistical Society **42**(2), 109–142 (1980)

88. McGeachie, M.: Msc.utility functions for ceteris paribus preferences. Master's thesis, Department of Electrical Engineering and Computer Science, MIT (2002)

89. McGeachie, M., Doyle, J.: Efficient utility functions for ceteris paribus preferences. In: Eighteenth national conference on Artificial intelligence, pp. 279–284. American Association for Artificial Intelligence, Menlo Park, CA, USA (2002)

90. McGeachie, M., Doyle, J.: Utility functions for ceteris paribus preferences. Computational Intelligence **20**(2), 158–217 (2002)

91. Meyer, P., Roubens, M.: Choice, Ranking and Sorting in Fuzzy Multiple Criteria Decision Aid. In: J. Figueira, S. Greco, M. Ehrgott (eds.) Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 471–506. Springer Verlag, Boston, Dordrecht, London (2005)

92. Miettinen, K.: Nonlinear Multiobjective Optimization, *International Series in Operations Research and Management Science*, vol. 12. Kluwer Academic Publishers, Dordrecht (1999)

93. Mousseau, V., Figueira, J., Naux, J.P.: Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. European Journal of Operational Research **130**(2), 263–275 (2001). DOI 10.1016/S0377-2217(00)00041-2

94. Olafsson, S., Li, X., Wu, S.: Operations research and data mining. European Journal of Operational Research **187**(3), 1429 – 1448 (2008). DOI 10.1016/j.ejor.2006.09.023

95. Oliveira, H.P., Magalhaes, A., Cardoso, M.J., Cardoso, J.S.: An accurate and interpretable model for bcct.core. In: Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6158–6161 (2010)

96. Ouyang, H., Gray, A.: Learning dissimilarities by ranking: from sdp to qp. In: International Conference on Machine Learning, pp. 728–735 (2008). DOI 10.1145/1390156.1390248

97. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11**(5), 341–356 (1982)

98. Pawlak, Z.: Rough set approach to knowledge-based decision support. European Journal of Operational Research **99**(1), 48–57 (1997). DOI 10.1016/S0377-2217(96)00382-7

99. Potharst, R., Bioch, J.C.: A decision tree algorithm for ordinal classification. In: Advances in Intelligent Data Analysis, pp. 187–198 (1999)

100. Potharst, R., Bioch, J.C.: Decision trees for ordinal classification. Intelligent Data Analysis **4**(2), 97–111 (2000)

101. Potharst, R., Feelders, A.J.: Classification trees for problems with monotonicity constraints. SIGKDD Explorations Newsletter **4**(1), 1–10 (2002). DOI http://doi.acm.org/10.1145/568574.568577

102. Presson, A., Yoon, N., Bagryanova, L., Mah, V., Alavi, M., Maresh, E., Rajasekaran, A., Goodglick, L., Chia, D., Horvath, S.: Protein expression based multimarker analysis of breast cancer samples. BMC Cancer **11**(1), 230 (2011). DOI 10.1186/1471-2407-11-230

103. Pyon, Y.S., Li, J.: Identifying gene signatures from cancer progression data using ordinal analysis. In: Bioinformatics and Biomedicine, 2009. BIBM '09. IEEE International Conference on, pp. 136 –141 (2009). DOI 10.1109/BIBM.2009.18

104. Rebelo, A., Tkaczuk, J., Sousa, R., Cardoso, J.: Metric learning for music symbol recognition (2011)

105. Rietveld, P., Ouwersloot, H.: Ordinal data in multicriteria decision making : a stochastic dominance approach to siting nuclear power plants. European journal of operational research **56**(2), 249–262 (1992)

106. Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C.S.: Quadratic programming feature selection. Journal of Machine Learning Research **11**, 1491––1516 (2010)

107. Roy, B.: The outranking approach and the foundations of electre methods. Theory and Decision **31**, 49–73 (1991). DOI 10.1007/BF00134132

108. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education (2003)

109. Saaty, T.L.: How to make a decision: The analytic hierarchy process. European Journal of Operational Research **48**(1), 9 – 26 (1990). DOI 10.1016/0377-2217(90)90057-I

110. Saaty, T.L., Vargas, L.G., Saaty, T.L., Vargas, L.G.: The seven pillars of the analytic hierarchy process. In: Models, Methods, Concepts & Applications of the Analytic Hierarchy Process, *International Series in Operations Research & Management Science*, vol. 34, pp. 27–46. Springer US (2001). DOI 10.1007/978-1-4615-1665-1_2

111. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: In NIPS. MIT Press (2004)

112. Seth, S., Príncipe, J.C.: Variable Selection: A Statistical Dependence Perspective. In: Proceeding of the Ninth International Conference on Machine Learning and Applications, pp. 931—936 (2010)

113. Shashua, A., Levin, A.: Ranking with large margin principle: Two approaches (2003)

114. Shen, L., Joshi, A.: Ranking and Reranking with Perceptron. Machine Learning **60**, 73–96 (2005)

115. Siskos, Y., Grigoroudis, E., Matsatsinis, N.: Uta methods. In: J. Figueira, S. Greco, M. Ehrgott (eds.) Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 297–344. Springer Verlag, Boston, Dordrecht, London (2005)

116. Siwik, L., Natanek, S.: Elitist evolutionary multi-agent system in solving noisy multi-objective optimization problems. Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on pp. 3319–3326 (2008)

117. Siwik, L., Natanek, S.: Solving constrained multi-criteria optimization tasks using Elitist Evolutionary Multi-Agent System. Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on pp. 3358–3365 (2008)

118. Smet, Y.D., Guzmán, L.M.: Towards multicriteria clustering: An extension of the k-means algorithm. European Journal of Operational Research **158**(2), 390 – 398 (2004). DOI 10.1016/j.ejor.2003.06.012. Methodological Foundations of Multi-Criteria Decision Making

119. Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, Wei Wang: Data mining curriculum: a proposal (Version 1.0) (2006). URL `www.kdd.org/curriculum/CURMay06.pdf`. Retrieved February 27, 2012

120. Sousa, R., Cardoso, J.S.: Ensemble of Decision Trees with Global Constraints for Ordinal Classification. In: 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011). Cordoba, Spain, Spain (2011)

121. Sousa, R., Oliveira, H.P., Cardoso, J.S.: Feature selection with complexity measure in a quadratic programming setting. In: Proceedings of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), pp. 524–531 (2011)

122. Spearman, C.: The proof and measurement of association between two things. American Journal of Psychology **15**, 72–101 (1904)

123. Sridhar, P., Madni, A., Jamshidi, M.: Multi-criteria decision making in sensor networks. Instrumentation Measurement Magazine, IEEE **11**(1), 24 –29 (2008). DOI 10.1109/IM-M. 2008.4449011

124. Sun, B.Y., Li, J., Wu, D., Zhang, X.M., Li, W.B.: Kernel discriminant learning for ordinal regression. Knowledge and Data Engineering, IEEE Transactions on **22**(6), 906 –910 (2010). DOI 10.1109/TKDE.2009.170

125. Tagliafico, A., Tagliafico, G., Tosto, S., Chiesa, F., Martinoli, C., Derchi, L.E., Calabrese, M.: Mammographic density estimation: Comparison among bi-rads categories, a semi-automated software and a fully automated one. The Breast **18**(1), 35–40 (2009)

126. Taha, H.A.: Operations Research: An Introduction (8th Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2006)

127. Tervonen, T., Figueira, J.R.: A survey on stochastic multicriteria acceptability analysis methods. Journal of Multi-Criteria Decision Analysis **15**, 1–14 (2008). DOI 10.1002/mcda.407

128. Tervonen, T., Lahdelma, R.: Implementing stochastic multicriteria acceptability analysis. European Journal of Operational Research **178**(2), 500–513 (2007). DOI 10.1016/j.ejor.2005. 12.037

129. Tutz, G.: Generalized Semiparametrically Structured Ordinal Models. Biometrics **59**, 263–273 (2003)

130. Ustinovichius, L., Zavadskas, E.K., Podvezko, V.: The application of a quantitative multiple criteria decision making (mcdm-1) approach to the analysis of investments in construction. Control and cybernetics **36** (2007)

131. Vanya, V.B., Kristiaan, P., K., S.J.A., Sabine, V.H.: Learning transformation models for ranking and survival analysis. Journal of machine learning research **12**, 819–862 (2011)

132. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)

133. Waegeman, W., Baets, B.D., Boullart, L.: Roc analysis in ordinal regression learning. Pattern Recognition Letters **29**(1), 1 – 9 (2008). DOI 10.1016/j.patrec.2007.07.019

134. Waegeman, W., De Baets, B., Boullart, L.: A comparison of different ROC measures for ordinal regression. In: Proceedings of the CML 2006 workshop on ROC Analysis in Machine Learning (2006)

135. Waegeman, W., De Baets, B., Boullart, L.: Kernel-based learning methods for preference aggregation. 4OR: A Quarterly Journal of Operations Research **7**, 169–189 (2009). DOI 10.1007/s10288-008-0085-5

136. Wang, J.J., Jing, Y.Y., Zhang, C.F.: Weighting methodologies in multi-criteria evaluations of combined heat and power systems. International Journal of Energy Research **33**(12), 1023–1039 (2009). DOI 10.1002/er.1527

137. Wang, J.J., Jing, Y.Y., Zhang, C.F., Zhao, J.H.: Review on multi-criteria decision analysis aid in sustainable energy decision-making. Renewable and Sustainable Energy Reviews **13**(9), 2263 – 2278 (2009). DOI 10.1016/j.rser.2009.06.021

138. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**, 207–244 (2009)

139. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: In Proc. 6th Online World Conference on Soft Computing in Industrial Applications, pp. 25–42 (2001)

140. Xia, F., Tao, Q., Wang, J., Zhang, W.: Recursive Feature Extraction for Ordinal Regression. Neural Networks, 2007. IJCNN 2007. International Joint Conference on pp. 78–83 (2007)
141. Xu, X., Zhou, C., Wang, Z.: Credit scoring algorithm based on link analysis ranking with support vector machine. Expert Syst. Appl. **36**, 2625–2632 (2009). DOI 10.1016/j.eswa. 2008.01.024
142. Yang, L., Jin, R.: Distance metric learning: A comprehensive survey. Tech. rep., Department of Computer Science and Engineering, Michigan State University (2006)
143. Yu, S., Yu, K., Tresp, V., Kriegel, H.P.: Collaborative ordinal regression. In: Proceedings of the 23rd international conference on Machine learning, ICML '06, pp. 1089–1096. ACM, New York, NY, USA (2006). DOI 10.1145/1143844.1143981
144. Zhang, Z., Kwok, J.T., Yeung, D.Y.: Parametric distance metric learning with label information. In: Proceedings of the 18th international joint conference on Artificial intelligence, pp. 1450–1452. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2003)
145. Zopounidis, C., Doumpos, M.: Building additive utilities for multi-group hierarchical discrimination: The M.H.DIS method. Optimization Methods and Software **14**(3), 219–240 (2000). DOI 10.1080/10556780008805801
146. Zopounidis, C., Doumpos, M.: Multicriteria classification and sorting methods: A literature review. European Journal of Operational Research **138**(2), 229–246 (2002). DOI 10.1016/ S0377-2217(01)00243-0
147. Zopounidis, C., Pardalos, P.M.: Handbook of multicriteria analysis. Applied Optimization 103. Berlin: Springer. xxv (2010). DOI 10.1007/978-3-540-92828-7