This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2563434, IEEE Transactions on Knowledge and Data Engineering

On Learning of Choice Models with Interactive Attributes

Manish Aggarwal
Indian Institute of Management Ahmedabad, India

Abstract

Introducing recent advances in the machine learning techniques to state-of-the-art discrete choice models, we develop an approach to infer the unique and complex decision making process of a decision-maker (DM), which is characterized by the DM's priorities and attitudinal character, along with the attributes interaction, to name a few. On the basis of exemplary preference information in the form of pairwise comparisons of alternatives, our method seeks to induce a DM's preference model in terms of the parameters of recent discrete choice models. To this end, we reduce our learning function to a constrained non-linear optimization problem. Our learning approach is a simple one that takes into consideration the interaction among the attributes along with the priorities and the unique attitudinal character of a DM. The experimental results on standard benchmark datasets suggest that our approach is not only intuitively appealing and easily interpretable but also competitive to state-of-the-art methods.

Index Terms—***Preference Learning; choice modelling; multi-attribute decision making; attitudinal character; attributes interaction***

1 Introduction

Multinomial logit (MNL) [2] is the easiest and most widely used discrete choice model based on the principle of utility maximization. The main reason for its popularity is the easy interpretability. The popularity of the MNL model can be gauged through its number of applications in the last two decades. It has been applied in severity analysis [24], [25], [34], price optimization [26], revenue optimization [42], location planning [27], choice analysis problems [29]–[31], [33], [47], risk analysis [28], [35], [38], [39], [43], [44], demand analysis [32], [36], data analytics [37], [40], [45], regression analysis [41], [48], [50], causal inference in medicine [46], and forecasting [49], to name a few.

Since the appearance of the MNL model, several extensions of the same have appeared in the literature. The nested logit [11], [12], GEV [13], multinomial probit (MNP) [14], paired combinatorial logit (PCL) [3], [4], cross-nested logit (CNL) [5], continuous CNL model [6], generalized nested logit (GNL) model [7], generalized MNL (GenMNL) model [8], mixed multinomial logit [9] and the fuzzy integral MNL model [10] are some of the popular variants of the MNL model. Recently, Choquet MNL (CMNL) and its extension attitudinal CMNL (ACMNL) are proposed in [57] to consider interaction among the attribute values. ACMNL model also considers the DM's attitudinal character, besides the interaction among the attributes.

The logit models bascially attempt to model the unique decision-making approach of a DM. In principle, we can infer the decision making model of a DM with the knowledge of the DM's choices, which can be used to predict the DM's choices for any set of alternatives. The problem of empirically inferring an individual's choice behaviour is a very interesting problem for economists, marketers, managers, and computer scientists alike. In this regard, seminal pieces of work have appeared in the literature. For instance, in [52], the attribute threshold values are investigated to determine whether the concerned alternatives stand a chance of being chosen or not. Some market response measures are calculated in [51] to infer the effect of marketing on consumer choices. An approach to preference segementation for identification of the determinants of brand switching and the consumers' response to price changes, is developed in [53].

These studies, however, do not make use of the preferences to learn a DM's predictive decision model that can be validated. In this regard, the work in [54] is quite relevant, in which the parameters of MNL model are estimated on the basis of rank ordered data of the form $\boldsymbol{a} \succ \boldsymbol{b} \succ \boldsymbol{c} \succ \ldots$, where $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ denote alternatives, and $\succ$ denotes the relation *preferred to*, for a DM. Guaranteeing the rank ordering, across the complete set of alternatives, in this learning approach is quite cumbersome, which limits its application in practice. More importantly, the crucial information regarding the interaction among the attributes, and the specific attitudinal character of a DM, is ignored in these works. Also, most of these works are based on data collected from survey questionnaire, and hence these models are neither fully verifiable, and automatized, nor very scalable either.

1.1 Motivation

We often find it easier to compare two options (on the basis of their desired attributes), and choose one, rather than assigning a quality score to all the available options, at once. Based on such pair-wise choices, it is possible to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2563434, IEEE Transactions on Knowledge and Data Engineering

understand an individual's decision behaviour, and predict his/her preferences. Predicting the preferences of others on the basis of their choices is a key part of social cognitive development [59]. Young children demonstrate the ability to make inferences about the preferences of others based on their choices. Decision making is a complex process that is specific to an individual. It involves different alternatives that are described by multiple attributes, and, many a times, there exists an interaction among the attributes, i.e. a few attributes, if present together in an alternative are much more (or less) valuable, than being individually. Besides, a DM's attitudinal character (tolerance) inevitably adds to this complexity. There are also inconsistencies in a DM's own decision behaviour, which add to the difficulty of infering the DM's choices.

We are motivated to somehow learn the unique decision behaviour of a DM, considering these features of human decision making. While, some of the existing methods have focussed on one of these features, our appraoach is distinguished by the fact that we consider all these practical aspects of human decision making, at the same time. In this regard, preference learning (PL) that is emerging as a new subfield of machine learning and data mining, appears very much interesting. PL is inspired from the human way of choosing the better of the two options, on the basis of desired attributes under consideration. We are concerned with the construction of preference model from a DM's preferences, with PL providing the technical background for *learning through preferences*, deeply rooted in our cognition. More specifically, we introduce recent algorithmic advances in PL to the area of choice modelling. While the choice models are concerned with an empirical analysis of a DM's choice behaviour through a quantitative approach, machine learning is a scientific discipline that explores different algorithms to learn from data. In this sense, choice models and machine learning can be argued to be complementary, which may aid each other. Traditionally, choice models put much emphasis on the interpretability and intuitive appeal. Machine learning, on the other side, is more focusing on computationally efficient algorithms for inducing predictive models, and prediction performance is given more importance than the interpretability.

Pointing to this interesting compelementarity, we feel interested to develop a preference learning approach to infer the parameters of the recent choice models. By this, we mean the idea of applying machine learning methods for a *preference-information* driven construction of state-of-the-art discrete choice models. More concretely, we apply PL techniques to learn the parameters of MNL, CMNL and ACMNL models. It would be worthwhile to mention that studies in [61]–[63] also deal with preferences-based learning, but with a different context. All these works are concerned with active learning in sequential decision making settings. They put an emphasis on selectively picking the training samples for the applications where the observations are time-consuming and/or expensive. In comparison, the present study considers all the training samples, at once, and is simpler by design.

1.2 The Proposed Work

We divide the sample of alternatives in multiple random preference pairs (of a DM) such that $\boldsymbol{a} \succ \boldsymbol{b}$. These preference tuples form the training information for learning a DM's preference model. In essence, we induce the behavioral model: Choose alternative $\boldsymbol{a}$ iff $U_{\boldsymbol{a}} > U_{\boldsymbol{b}}$. Both discrete choice models and PL are concerned with the construction of this kind of preference models of the form $\boldsymbol{a} \succ \boldsymbol{b}$. This commonality has inspired us to combine these two disciplines together to complement each other. By this, we mean the idea of applying PL-based machine learning methods for a *preference-information* driven construction of choice models. What we consider specifically interesting in this regard is the combination of the recent choice *modelling* methods and machine learning *algorithms* to infer a DM's choice behaviour from his/her preferences. Concretely, we use the *learning-to-rank* approach against the background of PL for the induction of recent choice models.

The paper is outlined as follows: Section 2 sets the background for the paper with an overview of the related concepts. Section 3 is concerned with modelling of the human decision process that is unique to an individual. In Section 4, we present a preferences-based approach to learn some predictive decision models. Section 5 gives the details of an experimental study to empirically validate the proposed learning approach. In Section 6, we actually learn the predictive models through an application of the proposed approach on some real datasets. Section 7 concludes the study.

2 Background

2.1 MNL Model

MNL model postulates that each alternative can be seen as a bundle of attributes. A DM makes choices among various alternatives so as to maximize the utility, i.e. choosing the alternative whose attributes collectively yield more utility than those of all other alternatives. Mathematically, it postulates that a DM derives from an alternative $\boldsymbol{a}_i$ a utility value U_i, given as:

$$U_i = V_i + \epsilon_i$$

where, ϵ_i is the unestimated utility, and V_i is the representative or systematic utility that is based on the observed attribute values and the attribute weight vector. If the value of m^{th} attribute of $\boldsymbol{a}_i$ is, say $a_i^{(m)}$, and $\beta^{(m)}$ is the corresponding weight that the DM attaches to m^{th} attribute. Then

$$V_i = \sum_{m=1}^{M} \beta^{(m)} a_i^{(m)} \tag{1}$$

where, M is the total number of attributes. It has been shown in [2] that probability P_i of an alternative $\boldsymbol{a}_i$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2563434, IEEE Transactions on Knowledge and Data Engineering

yielding the highest utility to the DM, and thus is chosen, is given by

$$\begin{aligned} P_i &= \frac{exp(V_i)}{\sum_{i=1}^{K} exp(V_i)} \\ &= \frac{exp(\sum_{m=1}^{M} \beta_i^{(m)} a_i^{(m)})}{\sum_{k=1}^{K} exp(\sum_{m=1}^{M} \beta_k^{(m)} a_k^{(m)})} \end{aligned} \tag{2}$$

The weight vector can be seen as the DM's *tastes* for the attributes, and it characterizes the DM's unique choice behaviour. The net utility derived from an alternative is computed as a function (weighted average) of the attribute values and the attribute weight vector. While MNL model reflects, to some extent, a DM's priorities (or the attribute weights), the individualistic characteristics such as the degree of interaction among the attributes, and a DM's attitudinal character[1] remain unrepresented. In order to address these drawbacks, MNL model is extended as Choquet MNL (CMNL) and attitudinal Choquet MNL (ACMNL) models in [57]. Before we delve upon CMNL and ACMNL models, we signify the characteristics of these models such as attributes interaction and a DM's attitudinal character in modelling the real world decision making.

2.2 Attributes Interaction

The conventional discrete choice models (such as MNL and its extensions) depend only on the attribute values and the attribute weight vector. In practice, however, there often exists some interaction among the attributes. That is, the attributes may not always be additive as assumed in the existing logit models of discrete choice. There may exist a synergy or redundancy among the attributes, and hence the importance (weight) depends upon the particular combination of attributes. Suppose, for instance, a programming task requires a mix of language and programming skills from the sets $A = \{\text{French, German, English}\}$, and $B = \{\text{Java, c++}\}$. Hence, the presence of the skills in combination from both A and B is more valued than their individual presence. If we represent the importance of the set of elements A as $\mu(A)$, then formally, a positive interaction can be expressed as: $\mu(A \cup B) > \mu(A) + \mu(B)$.

In a situation, when a combination of skills from the two sets A and B is utmost essential, then $\mu(A \cup B)$ can be high although $\mu(A) = \mu(B) = 0$, suggesting that the mere presence of either of the skills from A and B is unacceptable, as far as the task at hand is concerned. Likewise, a negative interaction may also exist among the attributes, when the attributes are redundant. For example, if $C = \{\text{J2EE}, \text{c\#}\}$, then B and C may be redundant, and their utilities diminish in a combination. That is, $\mu(B \cup C) < \mu(B) + \mu(C)$. These considerations motivate the use of the non-additive measures, also called as capacities or fuzzy measures [55] in the discrete choice models that assume that the attributes are always additive. In this regard, Choquet integral ($\mathcal{CI}$) [56] is an important aggregation function with the ability to represent interaction among the attributes.

1. The degree of conjunctiveness (and-ness) or disjunctiveness (or-ness) in the aggregation of the attribute values

2.3 Individual Attitudinal Character

Another drawback of the existing discrete choice models is the fact that it is very difficult to consider a DM's individual attitudinal character in arriving at the choice probabilities. The weighted mean, used in the discrete choice models, has no provision to consider the attitudinal character. That is, the aggregation result of the weighted averaging operator does not depend on the DM, and is only a function of the attribute values and the attribute weight vector. In the real world, however, the human aggregation is a complex process, and a DM's attitudinal character plays an important role in the same. A tolerant DM may be happy with just one of the attributes with an 'or-like' (disjunctive) behaviour, while his/her less tolerant counterpart may emphasize upon satisfying all the attributes with an 'and-like' (conjunctive) aggregation tendency. In other words, every individual displays a varying degree of compensation in the aggregation process. The more tolerant a DM is, the more is the compensation in the aggregation process, or it is more 'or-like' (disjunctive).

2.4 Choquet Multinomial Logit Models

Choquet multinomial logit (CMNL) model extends MNL model to cater to the interactive attributes. It is based on the concepts of game and capacity, defined on a vector $X = \{1, \ldots, M\}$. A game μ on X is a set function, $\mu : 2^X \to \mathbb{R}$ satisfying $\mu(\emptyset) = 0$. For any two $A, B \subseteq X$, a capacity (or fuzzy measure) μ on X is a game on X satisfying $\mu(A) \le \mu(B)$ whenever $A \subseteq B$. In particular, it follows that $\mu : 2^X \to [0, \infty)$. A capacity μ is *normalized*, when $\mu(X) = 1$. We consider a set of K alternatives characterized by M attributes.

Choquet multinomial logit (CMNL) model has the representative utility V_i given as a function of fuzzy measure μ, computed through $\mathcal{CI}$ operator, and is shown as

$$V_i = \mathcal{CI}_\mu = \sum_{m=1}^{M} a_i^{(\sigma(m))} (\mu(B^{(m)}) - \mu(B^{(m+1)})) \tag{3}$$

where, $\mu : 2^X \to [0, 1]$ is a fuzzy measure on a set of M attribute values of $\boldsymbol{a}_i$, $\sigma(m)$ indicates a permutation on M such that $a_i^{(\sigma(1))} \le \ldots \le a_i^{(\sigma(M))}$; $B^{(m)} = \{\sigma(m), \ldots, \sigma(M)\}$; and $B^{(M+1)} = \emptyset$. The form in (3) can be simplified using Möbius transform as follows:

$$\mathcal{CI}_\mu = \sum_{T \subseteq X} \mathbb{m}(T) \min \left\{ a_i^{(m)} \mid m \in T \right\} \tag{4}$$

where, $\min \left\{ a_i^{(m)} \mid m \in T \right\} := \min_{\{m \mid m \in T\}} \{a_i^{(m)}\}$, μ denotes fuzzy measure, and $\mathbb{m}_\mu$ refers to Möbius transform. For all $B \subseteq X$, Möbius transform $\mathbb{m}_\mu$ of measure μ is defined as follows:

$$\mathbb{m}_\mu(A) = \sum_{B \subseteq A} (-1)^{|A| - |B|} \mu(B) \tag{5}$$

The value $\mathbb{m}_\mu(A)$ can be interpreted as the weight that is *exclusively* allocated to the subset of attributes A, instead of being indirectly connected with A due to the interaction with other subsets. The choice probability in CMNL model is thus given as:

$$\begin{aligned} P_i &= \frac{exp\,(V_i)}{\sum_{k=1}^{K} exp(V_k)} \\ &= \frac{exp\left(\sum_{m=1}^{M} a_i^{(\sigma(m))}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\right)}{\sum_{k=1}^{K} exp\left(\sum_{m=1}^{M} a_k^{(\sigma(m))}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\right)} \\ &= \frac{exp\left(\sum_{T\subseteq X} \mathbb{m}(T) \min\left\{a_i^{(m)} \mid m \in T\right\}\right)}{\sum_{k=1}^{K} exp\left(\sum_{T\subseteq X} \mathbb{m}(T) \min\left\{a_k^{(m)} \mid m \in T\right\}\right)} \end{aligned} \tag{6}$$

where, $\min\left\{a^{(m)} \mid m \in T\right\} := \min\limits_{\{m\in T\}} \{a^{(m)}\}$.

2.5 Attitudinal Choquet Multinomial Logit Model

ACMNL model considers the varying attitudes of the DMs, along with the attributes interaction. It has the representative utility V_i computed as a function of an attitudinal parameter λ, fuzzy measure μ and the attribute values through $\mathcal{ACI}$ operator [58], shown as:

$$\begin{aligned} V_i &= \log_\lambda\left(\sum_{m=1}^{M}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\lambda^{a_i^{(\sigma(m))}}\right) \\ &= \sum_{T\subseteq X} \mathbb{m}(T)\lambda^{\min\left\{a_i^{(m)} \mid m\in T\right\}} \end{aligned} \tag{7}$$

where, $\lambda \in (0, \infty]$ indicates the DM's level of disjunctiveness. The more it is, the more is the output closer to $\max\{a^{(m)}\}_{m=1}^{M}$. Depending on the different DMs behavioural specifications represented by λ, one can have a range of ACMNL models. ACMNL model caters to the situations with interactive attribute values and the varying attitudinal characters of the DMs. Accordingly, the choice probability by ACMNL model is given as:

$$\begin{aligned} P_i &= \frac{exp\left(\log_\lambda\left(\sum_{m=1}^{M}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\lambda^{a_i^{(\sigma(m))}}\right)\right)}{\sum_{k=1}^{K} exp\left(\log_\lambda\left(\sum_{m=1}^{M}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\lambda^{a_k^{(\sigma(m))}}\right)\right)} \\ &= \frac{\sum_{m=1}^{M}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\lambda^{a_i^{(\sigma(m))}}}{\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right)\lambda^{a_k^{(\sigma(m))}}} \\ &= \frac{\sum_{T\subseteq X} \mathbb{m}(T)\lambda^{\min\left\{a_i^{(m)} \mid m\in T\right\}}}{\sum_{k=1}^{K}\sum_{T\subseteq X} \mathbb{m}(T)\lambda^{\min\left\{a_k^{(m)} \mid m\in T\right\}}} \end{aligned} \tag{8}$$

3 Modelling Human Decision Process

In this section, we highlight the advantages of $\mathcal{ACI}_{\mu,\lambda}$ operator in modelling a real world human decision. While one DM (AND-like) may stress on meeting all the attributes, another (OR-like) may be fine with only a few of the attributes meeting the expectations. We model a *decision making mind* in terms of the parameters μ and λ of $\mathcal{ACI}_{\mu,\lambda}$ operator, indicating the attributes interaction and attitudinal character, respectively. $\mathcal{ACI}_{\mu,\lambda}$ helps to simultaneously model a DM's level of conjunctiveness (or disjunctiveness) in the aggregation process along with the criteria interaction.

We consider a potential house-buyer who makes a choice considering a pair of interactive attributes. He would like to purchase a house with at least 4 rooms, and 2 garages. If an alternative has any of the attribute's value below the corresponding threshold, then this alternative would remain out of purview. We generate a set of 200 data points randomly, each indicating an available house for purchase, with attributes' values on a scale of $[1 : 8] \times [1 : 4]$.

In order to understand the effect of the buyer's attitudinal character on his decision behaviour, we generate decision boundaries for this buyer at different values of λ. The attribute values are normalized through linear scaling on an interval of $[0, 1]$, and the buyer's net evaluation for each of these alternatives is determined through $\mathcal{ACI}$ operator at $\lambda = 0.9, 1^+, 5, 10$, with a 2-fold cross validation. A graphical illustration of the buyer's decision boundaries is given in a 3-dimensional plot in Figure 1. The aggregation values shown are an average of the results obtained in 20 iterations. We take the normalized values of *number of rooms* and the *number of garages* on X and Y axes respectively, while giving the decision boundary on the Z-axis. If the evaluation score for an alternative is more than 0.5, then it would be a prospective *buy*, else not.

We note that the decision boundaries obtained at $\lambda = 5, 10$ are much more informative and complex than that obtained at $\lambda = 1^+$ ($\mathcal{CI}$ operator), which is reflective of a better consideration of the buyer's specific attitudinal character in the case of $\mathcal{ACI}$. To some extent it can also be explained by $\mathcal{ACI}$ operator's flexibility to rescale the original feature space, which helps in better modelling the utility that an attribute value holds for the DM. In comparison, $\mathcal{CI}$ operator, though models the criteria interaction, takes a neutral (constant) attitudinal character for each DM. We perform the same exercise using $\mathcal{CI}$, and the

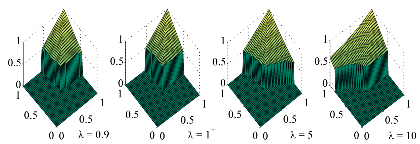

Fig. 1: The decision boundaries obtained with $ACI_{\mu,\lambda}$ operator at different λ values. X-axis indicates the number of rooms, while Y-axis indicates the number of garages. It depicts the decision boundaries obtained with 0.5 as a threshold for the attitudinal Choquet integral, i.e. when the output is greater than 0.5 the decision is 'buy' otherwise 'not buy'. As seen, for different values of λ, $\mathcal{ACI}$ serves more flexible decision boundary compared to the conventional Choquet integral ($\lambda = 1^+$).

decision boundary obtained is the same as that obtained with $\mathcal{ACI}$ at $\lambda = 1^+$. We observe that λ parameter in $\mathcal{ACI}$ leads to providing multiple evaluation scores for the same set of input attributes, thereby better modelling a DM's decision behaviour.

To the best of our knowledge, none of the existing operators take account of both the attributes interaction and the attitudinal character, at the same time. Perhaps this is the reason behind the conventional operators failing to give an aggregation output anyway close to the actual aggregated evaluation of attributes by the subjects in [64]–[66]. The flexibility on account of extra parameter λ and a consideration of attributes interaction through μ in $\mathcal{ACI}$ would be potentially of help in inferring the actual aggregation behaviour of the subjects. [2]

In the next section, we give an approach to empirically learn the parameters μ and λ from the given data. We verify the proposed approach in ascertaining the ability of $\mathcal{ACI}$ operator in modelling human decision by applying our approach to a set of datasets, in Section 5.

4 The Proposed Learning Model

4.1 The Basics of Our Learning Approach

In this section, we present a preferences-based approach to learn the parameters λ and μ of ACMNL model. In the real world, typically a DM faces a set of options (alternatives), and the DM chooses the alternative that yields the maximum utility to him by virtue of its attributes. For instance, a buyer chooses which product(s) to buy among several competing ones; an organization needs to determine the most profitable products to produce, best production technology, or the best supplier. The present work is about inferring a DM's predictive choice model. We give an outline of our work as follows:

- We observe a DM's pair-wise preferences of alternatives along with their corresponding attribute values.
- Through a collection of such preferences, we learn the parameters of MNL, CMNL and ACMNL models. Since, we deploy a PL-based learning methodology, we refer to them as PL-MNL, PL-CMNL, PL-ACMNL, respectively, in the sequel.
- Our learning objective is accomplished by using the learning-to-rank machine learning methods. Our model is probabilistic and, therefore, tolerant towards mistakes, incorrect preference statements, or variations in the DM's own attitudinal character.
- Having learned the DM's preference model, we predict a ranking for any new set of alternatives, a comparison of which with the ground-truth ranking offers a means to validate the performance of the proposed approach.
- We apply our learning approach on a set of 12 benchmark datasets. The prediction performances of PL-MNL, PL-CMNL, PL-ACMNL models are compared with those of state-of-the-art methods such as RankSVM, TOPSIS, and polynomial kernel as baselines.

2. For instance, the number of rooms (3 or 4) is instrumental in deciding the buyer's choice in this particular example, and leads to his decision (buy or not buy). From the buyer's point of view, there is a big difference in the utility derived from a house with 3 rooms and that with 4 rooms, or similarly between the houses with 1 and 2 garages.

We now present the settings of our learning model. We consider to be given a set of K alternatives, each of which is characterized by M attributes. An alternative $\boldsymbol{a}$ is represented as:

$$\boldsymbol{a} = \left(a^{(1)}, \ldots, a^{(M)}\right) \in \mathbb{R}^M,$$

where $a^{(m)}$ is the value of m^{th} attribute of $\boldsymbol{a}$. Each of the values, comprising vector $\boldsymbol{a}$ are normalized values, which can be obtained through either linear scaling of the attribute values in the interval $[0,1]$ (as shown in (37)), or through standardization appraoach (refer (38)). We consider that these values are monotone in the sense of the-higher-the-better, which leads to Pareto dominance relation:

$$\boldsymbol{a} \succ \boldsymbol{b} \text{ if } \forall m : a^{(m)} > b^{(m)} \tag{9}$$

where, $\boldsymbol{a} \succ \boldsymbol{b}$ implies that $\boldsymbol{a}$ is preferred to $\boldsymbol{b}$. However, very few instances may be found satisfying Pareto dominance, as shown in (9). Therefore, a refinement of this relation is sought to determine the total order (ordinal ranking) of the set of alternatives (or atleast the best alternative). Each DM has a unique model to determine the best alternative. Our objective is to learn the choice model of a DM from the DM's observed pair-wise choices.

To this end, we fit MNL, CMNL and ACMNL models to the observed preference data of the form $\boldsymbol{a} \succ \boldsymbol{b}$. That is, given the preference tuples of the form $(\boldsymbol{a}, \boldsymbol{b})$, we learn the parameters of MNL, CMNL and ACMNL models, specific to the DM. We randomly divide the set of alternatives $\mathcal{A}$ into two halves $\mathcal{A}_{train}$ and $\mathcal{A}_{test}$, such that $\mathcal{A}_{train}$ is used for training, and $\mathcal{A}_{test}$ is used as the testing dataset. From $\mathcal{A}_{train}$, we select a set $\mathcal{S} = \{s_1, \ldots, s_N\}$ of N pairwise preferences of the form:

$$s_n : \boldsymbol{a}_n \succ \boldsymbol{b}_n, \tag{10}$$

where $\boldsymbol{a}_n, \boldsymbol{b}_n \in \mathcal{A}_{train}$. The set $\mathcal{S} = \{s_1, \ldots, s_N\}$ constitutes the training information. We infer $\boldsymbol{\beta}$ in MNL model, λ in CMNL model and a pair of λ and $\boldsymbol{\mu} = \left(\mu(B^{(1)}), \ldots, \mu(B^{(M)})\right) \in \mathbb{R}^M$ in ACMNL model such that the preference relations of the form $s_n : \boldsymbol{a}_n \succ \boldsymbol{b}_n$ observed in $\mathcal{A}_{train}$ are preserved.

4.2 PL-MNL Model

MNL model characterizes a DM's choice behaviour through the attribute weight vector $\boldsymbol{\beta}$, specific to a DM. The unique choice model of a DM is specified by $\boldsymbol{\beta}$. Our learning model takes a DM's preference choices of the form $\boldsymbol{a} \succ \boldsymbol{b}$, as exemplary training information. Concretely, for the preference pair $(\boldsymbol{a}, \boldsymbol{b})$, our decision model is:

$$P(\boldsymbol{a} \succ \boldsymbol{b}|\boldsymbol{\beta}) = \frac{exp(V_{\boldsymbol{a}})}{exp(V_{\boldsymbol{a}}) + exp(V_{\boldsymbol{b}})}, \tag{11}$$

where, $V_{\boldsymbol{a}} = V(\boldsymbol{\beta}, \boldsymbol{a}) = \boldsymbol{\beta}.\boldsymbol{a} = \sum_{m=1}^{M} \beta^{(m)} a^{(m)}$. Here, $\{a^{(m)}\}_{m=1}^{M}$ values, comprising $\boldsymbol{a}$ are the normalized values. For the given preference pairs $\{s_n : \boldsymbol{a}_n \succ \boldsymbol{b}_n\}_n$, we learn $\boldsymbol{\beta}$ such that the relation $\boldsymbol{a} \succ \boldsymbol{b}$ is preserved.

Our learning model is bascially a variant of Bradley-Terry model [60], with a logistic (sigmoidal) link function. We specify the 'utility' of a decision alternative $\boldsymbol{a}$ as a function (weighted average) of the attributes and the corresponding weights, leading to a model of the form, shown in (11).

The advantages of this model is its simplicity, easy interpretability, and high computational efficiency. By this, we essentially mean that the probability of the preference $\boldsymbol{a} \succ \boldsymbol{b}$ increases with increasing attribute values of $\boldsymbol{a}$ and decreasing attribute values of $\boldsymbol{b}$. In the extreme cases where $V_{\boldsymbol{a}} \to 0$ or $V_{\boldsymbol{b}} \to \infty$, the probability of choosing $\boldsymbol{a}$ goes to 0. Similarly, if $V_{\boldsymbol{a}} = V_{\boldsymbol{b}}$, the DM is supposed to choose one of the two randomly.

We accomplish our objective of learning $\boldsymbol{\beta}$ through the maximum likelihood estimation (MLE). The probability of the complete sample $\mathcal{S}$ is given by:

$$\begin{aligned} P(\mathcal{S}|\boldsymbol{\beta}) &= \prod_{n=1}^{N} P(\boldsymbol{a}_n \succ \boldsymbol{b}_n|\boldsymbol{\beta}) \\ &= \prod_{n=1}^{N} \frac{exp\left(\sum_{m=1}^{M} \beta^{(m)} a_n^{(m)}\right)}{exp\left(\sum_{m=1}^{M} \beta^{(m)} a_n^{(m)}\right) + exp\left(\sum_{m=1}^{M} \beta^{(m)} b_n^{(m)}\right)} \end{aligned} \tag{12}$$

This probability serves as a point of departure for the estimation of $\boldsymbol{\beta}$. The likelihood function for $\boldsymbol{\beta}$, obtained with MLE inference is given by

$$L(\boldsymbol{\beta}) = P(\mathcal{S}|\boldsymbol{\beta}). \tag{13}$$

The maximum likelihood estimator $\boldsymbol{\beta}^*$ maximizes the likelihood function $L(\boldsymbol{\beta})$ such that

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}_+^M} L(\boldsymbol{\beta}) \tag{14}$$

Since, log function is monotonically related to its arguments, we take the logarithm of the likelihood function in (13) for the computational convenience. Towards finding the ML estimator $\boldsymbol{\beta}^*$, we minimize the negative logarithm. Thus our learning problem is essentially an optimization problem that is given as:

$$\begin{aligned} -\ell(\boldsymbol{\beta}) &= -\log(L(\beta)) = -\sum_{n=1}^{N} V_{\boldsymbol{a}_n} + \sum_{n=1}^{N} \log\left(exp\left(V_{\boldsymbol{a}_n}\right) + exp\left(V_{\boldsymbol{b}_n}\right)\right) \longrightarrow \min \\ \Longrightarrow -\ell(\boldsymbol{\beta}) &= \\ & -\sum_{n=1}^{N}\sum_{m=1}^{M} \beta^{(m)} a_n^{(m)} + \sum_{n=1}^{N} \log\left(exp\left(\sum_{m=1}^{M} \beta^{(m)} a_n^{(m)}\right) + exp\left(\sum_{m=1}^{M} \beta^{(m)} b_n^{(m)}\right)\right) \longrightarrow \min \end{aligned} \tag{15}$$

4.3 PL-CMNL Model

In practice, often there exists a positive or negative interaction among the attributes, and a DM's choice behaviour is shaped by the attributes interaction. CMNL model explicitly considers the attributes interaction through $\mathcal{CI}$ operator. We recall from (3) and (4) that $V_{\boldsymbol{a}}$ in CMNL model is given as:

$$V_{\boldsymbol{a}} = \mathcal{CI}_{\boldsymbol{\mu}}(\boldsymbol{a}) = \sum_{T \subseteq X} \mathrm{m}(T) \min\left\{a^{(m)} \mid m \in T\right\} \tag{16}$$

We consider a DM's latent utility function $u(\cdot)$ that specifically determines the DM's choice behaviour, and shown as:

$$u(\boldsymbol{a}) = \mathcal{CI}_{\boldsymbol{\mu}}(\boldsymbol{a}) \tag{17}$$

Given a set of pairs $\mathcal{P} = \left\{(\boldsymbol{a}_n, \boldsymbol{b}_n) \mid l_n^a > l_n^b, 1 \le n \le N\right\}$, where $l_n^{\boldsymbol{a}}$ and $l_n^{\boldsymbol{b}}$ are the labels for $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$ respectively, we learn the vector $\mathcal{M} = \{\mathrm{m}(T)\}, \forall T \in X$ of $\mathcal{CI}_{\boldsymbol{\mu}}$ operator such that

$$\sum_{T \subseteq X} \mathrm{m}(T) \min\left\{a^{(m)} \mid m \in T\right\} > \sum_{T \subseteq X} \mathrm{m}(T) \min\left\{b^{(m)} \mid m \in T\right\} \tag{18}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2563434, IEEE Transactions on Knowledge and Data Engineering

We accomplish this objective by using the empirical risk minimization approach to find the optimal parameters. More concretely, we maximize the margin between the preferred alternatives to the non-preferred ones.

Let $\mathscr{M}$ denotes the margin (to be maximized), $\xi_i^{\boldsymbol{a}}$'s and $\xi_i^{\boldsymbol{b}}$'s be the slack variables related to soft margin, and γ is a trade-off parameter that controls the flexibility of the model, i.e. the higher the γ value, the more are the slacks punished. Then our constrained optimization problem to be solved can be formalized as follows:

$$\ell(\mathcal{M}) = \max_{\mathscr{M},\xi_1,\ldots,\xi_N} \left\{ \mathscr{M} - \frac{\gamma}{|\mathcal{P}|} \sum_{(\boldsymbol{a}_n,\boldsymbol{b}_n)\in\mathcal{P}} \left(\xi_n^{\boldsymbol{a}} + \xi_n^{\boldsymbol{b}}\right) \right\} \tag{19}$$

s.t.

$$\mathcal{CI}_{\boldsymbol{\mu}}(\boldsymbol{a}) - \mathcal{CI}_{\boldsymbol{\mu}}(\boldsymbol{b}) > \mathscr{M} - \xi_n^{\boldsymbol{a}} - \xi_n^{\boldsymbol{b}} \quad \forall(\boldsymbol{a}_n,\boldsymbol{b}_n)\in\mathcal{P} \tag{20}$$

$$\xi_n^{\boldsymbol{a}} \ge 0,\;\; \xi_n^{\boldsymbol{b}} \ge 0 \qquad \forall n\in\{1,\ldots,N\} \tag{21}$$

$$\sum_{T\subseteq X} \mathbb{m}(T) = 1 \tag{22}$$

$$\sum_{B\subseteq A} \mathbb{m}(B) \ge 0 \qquad \forall A\subseteq X \tag{23}$$

$$\sum_{L\subseteq A} \mathbb{m}(L) \le \sum_{\mathscr{L}\subseteq B} \mathbb{m}(\mathscr{L}) \qquad \forall A\subset B\subseteq X \tag{24}$$

4.4 PL-ACMNL Model

ACMNL model extends CMNL model with a consideration of a DM's attitudinal character, besides the attributes interaction. We recall from (7) that the representative utility is given as:

$$\begin{aligned} V_{\boldsymbol{a}} &= \log_\lambda \left(\sum_{m=1}^{M} \left(\mu(B^{(m)}) - \mu(B^{(m+1)})\right) \lambda^{a^{(\sigma(m))}} \right) \\ &= \sum_{T\subseteq X} \mathbb{m}(T) \lambda^{\min\left\{a^{(m)}|m\in T\right\}} \end{aligned} \tag{25}$$

where, the DM's choice probability depends on $\boldsymbol{\mu}$ and λ, both of which are specific to the DM.

Here, we model $u(.)$ through $\mathcal{ACI}$ operator. That is:

$$u(\boldsymbol{a}) = \mathcal{ACI}_{\boldsymbol{\mu},\lambda}(\boldsymbol{a}) \tag{26}$$

For a given $\mathcal{P}$, we learn the parameters $\boldsymbol{\mu}$ and λ of $\mathcal{ACI}_{\boldsymbol{\mu},\lambda}$ operator such that

$$\sum_{T\subseteq X} \mathbb{m}(T)\lambda^{\min\left\{a^{(m)}|m\in T\right\}} > \sum_{T\subseteq X} \mathbb{m}(T)\lambda^{\min\left\{b^{(m)}|m\in T\right\}} \tag{27}$$

We use the empirical risk minimization approach, and maximize the margin between the preferred alternatives to the non-preferred ones. The constrained optimization problem is formalized as follows:

$$\ell(\mathcal{M}) = \max_{\mathscr{M},\xi_1,\ldots,\xi_N} \left\{ \mathscr{M} - \frac{\gamma}{|\mathcal{P}|} \sum_{(\boldsymbol{a}_n,\boldsymbol{b}_n)\in\mathcal{P}} \left(\xi_n^{\boldsymbol{a}} + \xi_n^{\boldsymbol{b}}\right) \right\} \tag{28}$$

s.t.

$$\mathcal{ACI}_{\boldsymbol{\mu},\lambda}(\boldsymbol{a}) - \mathcal{ACI}_{\boldsymbol{\mu},\lambda}(\boldsymbol{b}) > \mathscr{M} - \xi_n^{\boldsymbol{a}} - \xi_n^{\boldsymbol{b}} \quad \forall(\boldsymbol{a}_n,\boldsymbol{b}_n)\in\mathcal{P} \tag{29}$$

$$\xi_n^{\boldsymbol{a}} \ge 0,\;\; \xi_n^{\boldsymbol{b}} \ge 0 \qquad \forall n\in\{1,\ldots,N\} \tag{30}$$

$$\lambda > 0,\;\; \lambda \neq 1 \tag{31}$$

$$\sum_{T\subseteq X} \mathbb{m}(T) = 1 \tag{32}$$

$$\sum_{B\subseteq A} \mathbb{m}(B) \ge 0 \qquad \forall A\subseteq X \tag{33}$$

$$\sum_{L\subseteq A} \mathbb{m}(L) \le \sum_{\mathscr{L}\subseteq B} \mathbb{m}(\mathscr{L}) \qquad \forall A\subset B\subseteq X \tag{34}$$

4.5 Regularization

In order to prevent overfitting, we regularize our likelihood function by introducing an additional parameter $\alpha \ge 0$. The regularization method is a standard means for controlling the capacity of a model and, therefore, to avoid a possible over-fitting of the data. We observe that maximizing regularized version of our likelihood functions proved to be useful. It also helped to deal with the cases where MLE did not return a unique maxima by restricting the norm of the parameter vector.

A common regularization approach in machine learning is to minimize objective functions of the form

$$f(\theta) = e(\theta) + \alpha.c(\theta). \tag{35}$$

where θ is a parameter vector specifying the learning model, $e(\theta)$ is the loss or error function giving a measure of the fitment of the model specified by θ with the data, $c(\theta)$ denotes a measure of the model capacity, (.) is either the L1 norm or the squared L2 norm, and α is a free parameter that needs to be tuned empirically (typically by cross-validation).

The regularization parameter α is introduced to achieve a trade-off between generalization and the accuracy of the learning model, i.e. to accurately reproduce the training data while avoiding overly complex models. In our case, the model error is given as the (negative) likelihood function ($e(.) = -\ell(.)$). Out of L1 and L2 norms, the use of standard L2 regularization turned out to be better. Thus our regularized objective function to be minimized is shown as:

$$f(.) = -\ell(.) + \alpha\left\|.\right\|^2 \;. \tag{36}$$

For the implementation of the same, we use *fmincon* function in the optimization toolbox of MATLAB.

5 Experimental Study

We test the validity of our learning approaches through an experimental study in the context of learning-to-rank. We compare the results obtained with PL-MNL, PL-CMNL, and PL-ACMNL models, with three standard and popular learning-to-rank approaches as baselines, namely, ranking support vector machine (RankSVM) [17], TOPSIS (short for 'technique for order preference by similarity to ideal solution') [16], and polynomial kernel logistic regression.

5.1 Data

Since the proposed method positions itself at the cross-junction of choice modelling and machine learning, its evaluation requires the datasets meeting the requirements of both these disciplines. The learning-to-rank approach that is commonly used in both machine learning as well as choice models facilitates the empirical validation of our model. For the task, we require the data in the form of multiple instances, with each instance a vector of multi-attribute values, along with the corresponding ordinal ranking by the DM. More importantly, the attribute values should have a monotone influence on the ranking of an instance. The real-world datasets meeting these twin requirements are quite difficult to find.

Fortunately, we could make use of the same datasets that have been considered in a learning model in [18]. The datasets give the ordinal rankings for alternatives and are monotone. The datasets could be found in UCI repository [19] and Weka machine learning framework [20]. Besides, the real world datasets about the evaluation of mathematical journals [21] and houses in the city of Den Bosch [22] are also used. We give a summary of the datasets in Table 1. For a detailed description of the datasets, we refer to [18].

Each of these datasets are characterized with the ordinal class information that is used as the ground truth in our model. By ordinal class, we mean that for an alternative $\boldsymbol{a}_k$, the actual rank r_k is a label (class) $l(\boldsymbol{a}_k) \in \mathcal{L} = \{l_1, l_2, \ldots, l_E\}$. The labels (for example, $1, 2, \ldots,$) are such that it is possible to sort them as $l_1 \leq l_2 \leq \ldots \leq l_E$ from the worst l_1 to the best l_E. The presence of labels helps to generate the pairwise preferences in our experiments.

As a preprocessing step, we normalize the data to ensure commensurability in the attribute values. The data is normalized by using two approaches. The first one is the linear scaling of the attribute values to the unit interval. Each attribute value $a^{(m)}, m = 1, \ldots, M$ is scaled as

$$a^{(m)} = \frac{a^{(m)} - L^{(m)}}{U^{(m)} - L^{(m)}} \in [0, 1] \tag{37}$$

where, $L^{(m)}$ and $U^{(m)}$ denote, respectively, the largest and smallest values among the attribute values $a^{(m)}, \forall \boldsymbol{a} \in \mathcal{A}$. In the second approach that is quite common in statistics, we standardize the data to have zero mean and unit variance. To this end, $a^{(m)}$ is replaced by

$$a^{(m)} = \frac{a^{(m)} - \mu^{(m)}}{\sigma^{(m)}} \tag{38}$$

where, $\mu^{(m)}$ and $\sigma^{(m)}$ denote, respectively, the mean and standard deviation in the attribute values $a^{(m)}, \forall \boldsymbol{a}$. For the purpose of our experimental study, we tried both normalized and standardized data, and found the latter to be more suitable.

5.2 Baseline and the Proposed Methods

RankSVM: RankSVM [17] is a state-of-the-art method in the field of learning-to-rank. Similar to the proposed PL-CMNL and PL-ACMNL models, it is a pairwise method based upon large margin maximization. The central principle in this method is to minimize the regularized margin-based pairwise loss. As in the case of our ranking model, RankSVM model is also constructed from a set of preference pairs of the form $s : (\boldsymbol{a} \succ \boldsymbol{b}) \in \mathcal{S}$, where $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{A}_{train}$. Once, such preference pairs are obtained, a scoring function $u(.)$ is induced from the preference pairs $s_n, n = 1, \ldots, N$, preserving the preference relations $s_n : \boldsymbol{a}_n \succ \boldsymbol{b}_n|_{n=1}^{N}$. That is,

$$\forall(\boldsymbol{a}_n, \boldsymbol{b}_n) \in \mathcal{S} : u(\boldsymbol{a}_n) > u(\boldsymbol{b}_n) \tag{39}$$

We implement rankSVM model by minimizing the objective function:

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{\beta}\|^2 \; + \gamma \sum_{s_n \in \mathcal{S}} c\xi_n \\ &\qquad s.t. : \boldsymbol{\beta}|\boldsymbol{a}_n - \boldsymbol{b}_n| \geq 1 - \xi_n, \;\; \forall\{s_n : (\boldsymbol{a}_n, \boldsymbol{b}_n)\} \\ &\qquad \xi_n \; \geq \; 0 \end{aligned} \tag{40}$$

where, $\|\boldsymbol{\beta}\|^2$ contributes to maximizing the margin, c is a constant, ξ_n is a slack variable corresponding to preference pair s_n, and γ is the tradeoff parameter between the margin size and magnitude of error. The pairwise difference vectors $|\boldsymbol{a}_n - \boldsymbol{b}_n|$ constitute the support vectors that satisfy the constraints, as given in (40). We have chosen RankSVM model as a baseline, mainly because of its popularity as a preferred preference model, high prediction performance, and the commonalities with the proposed model.

The proposed PL-based learning models and RankSVM both use the pairwise preferences as the training information. Secondly, while RankSVM is based upon margin maximization adding to the generalization in the learning process, PL-MNL, PL-CMNL and PL-ACMNL are probabilistic in nature based upon maximizing the probability for $(\boldsymbol{a} \succ \boldsymbol{b})$. More specifically, we have used the spider implementation[3] of RankSVM with a linear kernel. An important parameter in the RankSVM model is the regularization parameter that is selected from $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$ by means of an internal 5-fold cross validation on the training data.

TOPSIS: Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [16] is a state-of-the-art method from multi-attribute decison making (MADM) area, and is used to solve the choice and ranking problems. It helps to arrive at a total order (ranking) for a set of alternatives $\mathcal{A}$. In this method, one derives the positive ideal solution (PIS) and a negative ideal solution (NIS)

3. http://people.kyb.tuebingen.mpg.de/spider/

TABLE 1: Data sets and their properties

data set	#instances	#attributes	#classes	source
Scientific Journals (SJ)	172	5	4	[21]
CPU	209	6	4	UCI
Employee Selection (EMP)	488	4	4	WEKA
Mamographic (MAMMO)	830	5	2	UCI
Lecturers Evaluation (LEC)	1000	4	5	WEKA
Car Evaluation (CAR)	1728	6	4	UCI
Den Bosch (DB)	120	8	2	[22]
BreastCancer (BC)	286	9	2	UCI
Employee Acc/Rej (EMP)	1000	4	9	WEKA
Auto MPG (AUTO)	398	8	6	UCI
Social Workers (SW)	1000	8	5	WEKA
Concrete Compressive Strength (CONC)	1030	8	6	UCI

such that the benefit attributes attain the maximum values (among all the respective evaluations of various alternatives) while the cost attributes attain the minimum in PIS, and the vice-versa in NIS. The various alternatives are ranked such that the best alternative has the shortest distance from PIS and the farthest distance from NIS concurrently.

In our implementation of TOPSIS method, we perform the following steps.

1) The attributes values from $\mathcal{A}_{train}$ are standardized as per (38).
2) Using the normalized data, we *learn* the ideal solutions for $\mathcal{A}_{train}$. By *learning*, we mean that we look for the maximum and minimum values in $\mathcal{A}_{train}$. For example, if we are having 5 attributes, and the maximum values for these 5 attributes across all the alternatives in $\mathcal{A}_{train}$ are $1, 1, 0.90, 1, 0.95$, then PIS would be $(1, 1, 0.90, 1, 0.95)$. Naturally, the learnt PIS and NIS are different in each iteration depending upon the random selection of alternatives in $\mathcal{A}_{train}$ and $\mathcal{A}_{test}$. Similarly, NIS is computed with minimum values in $\mathcal{A}_{train}$.
3) The distance of alternative $\boldsymbol{a}_i$ in $\mathcal{A}_{test}$ are computed from the learnt PIS and NIS, and are ranked in the increasing order of the closeness coefficient, given as

$$D_i = \frac{\|\boldsymbol{a}_i - \boldsymbol{a}_\oplus\|}{\|\boldsymbol{a}_i - \boldsymbol{a}_\oplus\| + \|\boldsymbol{a}_i - \boldsymbol{a}_\ominus\|}$$

where, $\|.\|$ denotes the 2-norm distance, and $\boldsymbol{a}_\oplus$ and $\boldsymbol{a}_\ominus$ represent PIS and NIS, respectively. D_i acts as a kind of scoring function $u(\boldsymbol{a}_i)$.

Polynomial Kernel: The proposed learning approach, based on CMNL and ACMNL models can be seen as an extension of the conventional logistic regression, with the difference that while the logistic regression is *linear* in the input attributes, CMNL and ACMNL models are more flexible with their ability to model the non-linear dependencies between the attributes, as well. Hence, in order to study the increased flexibility in the proposed approach, we consider as a baseline, the logistic regression with polynomial kernel (POLY-kernel) that can also model non-linear dependencies between variables. The degree of the POLY-kernel is taken as 2, so as to consider low level interactions too.

RBF-kernel: Since, radial basis function (RBF) is able to capture interactions of the higher order through a Gaussian function, and is very flexible, it is a natural choice as a baseline. Like POLY-kernel, RBF also captures the non-linear dependencies in the input attributes. The RBF-kernel function is given as:

$$K(\boldsymbol{x}, \boldsymbol{x}^*) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}^*\|^2}{2\sigma^2}\right) . \quad (41)$$

The parameter $\sigma \in \mathbb{R}_+$ can drastically change the kernel characteristics. For a large value of σ, RBF-kernel behaves almost linearly, whereas for the smaller values, it yields a non-linear decision boundary. There is a considerable influence of σ on the overall model, and typically it is adjusted before the learning process. In our implementation, we use a nested cross validation and have chosen the optimal values among $\Sigma = \{10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$.

PL-MNL: Our model takes as input a subset of alternatives (for example, a set of all books or movies), $\mathcal{A}_{train} \to \mathcal{A}$, and produces as output the decision making model of the DM in the form of the $\boldsymbol{\beta}$ vector that can be deployed to predict the ranking (total order) of any set of new alternatives, preserving the preference orders:

$$\begin{aligned} &\boldsymbol{a} \succeq \boldsymbol{b} \Leftrightarrow \boldsymbol{\beta}.\boldsymbol{a} \geq \boldsymbol{\beta}.\boldsymbol{b} \\ &\quad\implies u(\boldsymbol{a}) \geq u(\boldsymbol{b}), \qquad \forall s : (\boldsymbol{a}, \boldsymbol{b}) \in \mathcal{S} \end{aligned}$$

Here, $\boldsymbol{\beta}$ performs the same role as the scoring function $u(.)$.

PL-CMNL and PL-ACMNL: Our model takes as input a subset of alternatives (for example, a set of all books or movies), $\mathcal{A} \to \mathcal{A}_{train}$, and gives the DM's decision model specified by $\boldsymbol{\mu}$ and λ, which can be deployed to predict the ranking (total order) for any set of new alternatives, preserving the preference orders:

$$\boldsymbol{a} \succeq \boldsymbol{b} \Longleftrightarrow \sum_{T \subseteq X} \mathrm{m}(T) \min\left\{a_i^{(m)} \mid m \in T\right\} > \sum_{T \subseteq X} \mathrm{m}(T) \min\left\{b_i^{(m)} \mid m \in T\right\}$$

$$\implies u(\boldsymbol{a}) \geq u(\boldsymbol{b}), \forall s : (\boldsymbol{a}, \boldsymbol{b}) \in \mathcal{S} \quad \text{in PL-CMNL model and,}$$

$$\boldsymbol{a} \succeq \boldsymbol{b} \Longleftrightarrow \sum_{T \subseteq X} \mathrm{m}(T) \lambda^{\min\left\{a_i^{(m)} | m \in T\right\}} > \sum_{T \subseteq X} \mathrm{m}(T) \lambda^{\min\left\{b_i^{(m)} | m \in T\right\}}$$
$$\Longrightarrow u(\boldsymbol{a}) \geq u(\boldsymbol{b}), \forall s : (\boldsymbol{a}, \boldsymbol{b}) \in \mathcal{S} \quad \text{in PL-ACMNL model}$$

The scoring function $u(.)$ is specified by $\mathrm{m}(T)$ in CMNL model and $(\mathrm{m}(T), \lambda)$ in ACMNL model.

6 Empirical Evaluation

6.1 Experiment Steps

We implement TOPSIS, RankSVM, PL-MNL, PL-CMNL, and PL-ACMNL methods on a set of 12 datasets shown in Table 1 by performing the following steps:

1) First, the data is randomly divided into two halves $\mathcal{A}_{train}$ and $\mathcal{A}_{test}$, for training and testing.
2) From $\mathcal{A}_{train}$, $N = 500$ preference pairs of the form $\mathcal{S} = \{s_n : (\boldsymbol{a}_n \succ \boldsymbol{b}_n)\}_{n=1}^{N=500}$ are generated through random sampling, which constitute the training information. For a fair comparison, it is ensured that $\mathcal{S}$ remains the same across the methods, in an iteration.
3) The learning models are induced on the set $\mathcal{S}$.
4) On the basis of the corresponding scoring function learnt for each of the methods, the ranking of alternatives in $\mathcal{A}_{test}$ is predicted.
5) The performance accuracy value for each of the methods is arrived at by comparing the predicted ranking with the ground truth ranking through C-index [23] (discussed next).
6) This whole procedure is repeated 100 times.
7) The average accuracy value along with the respective standard deviation is reported in Table 2.

6.2 Performance Evaluation

The ranking of an alternative $\boldsymbol{a} \in \mathcal{A}_{test}$ is predicted based upon the corresponding $u(\boldsymbol{a})$, induced for each of the methods. The performance of a method is evaluated by comparing the predicted rankings for the alternatives in $\mathcal{A}_{test}$ with the ground truth rankings. The extent of this agreement would give the performance measure for a method. Since, we have the ground truth for $\mathcal{A}_{test}$ in the form of labels (classes) that can be linearly ordered, an ordered partition of $\mathcal{A}_{test}$ can be generated as $\underline{\mathcal{A}}_{test} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_E\}$, where $\mathcal{A}_i = \{\boldsymbol{a}_k \in \mathcal{A}_{test} \,|\, r_k = l_i\}$.

Based upon the concordance of actual and predicted rankings, a performance measure termed as C-index is proposed in [23] to handle such tasks. Mathematically, C-index is defined as

$$C(u,\, \underline{\mathcal{A}}_{test}) = \frac{\sum_{1 \leq i < j \leq k} \sum_{(\boldsymbol{b}, \boldsymbol{a}) \in \mathcal{A}_i \times \mathcal{A}_j} F(u(\boldsymbol{b}),\, u(\boldsymbol{a}))}{\sum_{i<j} |\mathcal{A}_i| \,.\, |\mathcal{A}_j|} \tag{42}$$

where, $|.|$ gives the cardinality of a set and

$$F(d, e) = \begin{cases} 1 & d > e \\ 0 & otherwise \end{cases} \tag{43}$$

indicates whether or not a pair of alternatives has the correct preference order. Thus, for a preference tuple $(\boldsymbol{a} \succ \boldsymbol{b}) \in \underline{\mathcal{A}}_{test}$, C-index checks if $u(\boldsymbol{a}) > u(\boldsymbol{b})$, i.e. whether the model correctly suggests that $\boldsymbol{a} \succ \boldsymbol{b}$ on the basis of the learnt scoring function $u(.)$. C-index yields a fraction of correct pairwise comparisons of this kind. It would be worthwhile to mention that in the case of binary classification, C-index reduces to the area under the receiver operating characteristic (ROC) curve.

The steps carried out to arrive at C-index accuracy are summarized as follows:

1) A total ranking of the alternatives, $\{\boldsymbol{a}\} \in \mathcal{A}_{test}$, along with their respective values for $\{u(\boldsymbol{a})\}$ is obtained by sorting the alternatives as per the decreasing order of the ground truth ranking for $\mathcal{A}_{test}$.
2) The number of correct orderings such that $u(\boldsymbol{a}) > u(\boldsymbol{b})$ forms the numerator of the relation in (42).
3) The accuracy value is obtained by dividing the count of correct comparisons by the total number of comparisons, as shown in (42).

6.3 Results

The accuracy results, in terms of the C-index accuracy, are summarized in Table 2. The accuracy values reported are the average of the 100 accuracy values along with the standard deviation, obtained for each method and a dataset. Against each accuracy value, the corresponding rank is mentioned in parenthesis, with 1 as the best, and 5 as the worst. The overall performance of a method is arrived at by taking an average of these ranks, which is given in the last row. We observe that the proposed approach performs quite well in accurately predicting the choice probabilities, and the prediction performance is competitive with the state-of-the-art methods.

Overall, PL-ACMNL and PL-CMNL methods methods seem to be the outperformers. While CMNL model is equipped to model the interaction among the attributes, ACMNL model is even more flexible with its consideration of the DM's attitudinal character also. Clearly, PL-ACMNL is a winner among all, perhaps due to its flexibility. Theoretically, the better results obtained are on account of the ability of the underlying CMNL and ACMNL models to capture the non-linear dependencies in the input attributes. The same is true for POLY-kernel too, and hence the performance of POLY-kernel and PL-CMNL is found to be comparable. In contrast, PL-MNL model has not performed as well as its non-linear counterparts, perhaps due to it being essentially a linear model, and concomitantly its inability to model the interaction among the attributes. The effect of increase of an attribute is always the same in MNL model, regardless of the presence of other attributes, which does not model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2563434, IEEE Transactions on Knowledge and Data Engineering

the real world aggregation processes, as well. The DM's attitudinal character also remains unconsidered.

We also observe that rank-SVM has performed fairly well, and better than TOPSIS, and PL-MNL, which matches well with the expectations. Rank-SVM is a quite complex model capable of better modelling the relationships between input and output variables than TOPSIS and PL-MNL models. Both TOPSIS and PL-MNL models remain very much sensitive to any noisy value that can cause the learning process to go completely erratic. We also note that RBF-kernel's overall performance is not as good as its counterpart POLY-kernel. It indicates that in the given datasets, the low level interactions are having much more influence on the output variable than the high level interactions, and that could be the reason of outperformance of POLY-kernel over RBF-kernel.

On this note, we would also like to emphasize that no one approach could be said to be the best approach, as every learning approach has its own pros and cons. For instance, while, a complex learning approach such as rank-SVM or PL-ACMNL may be better in a few situations, it comes at an increased computational and implementation cost, which in a few situations may not be desirable. In addition, these approaches may not be as easily interpretable as TOPSIS and MNL. Lastly, and more importantly, it also depends on the type of data. In a dataset, with not much dependencies among the attributes, the simpler approaches such as TOPSIS and MNL may perform equally or even better than their complex counterparts. The idea of this study is to show the performance of different models on different datasets, and thereby to compare the performance of PL-CMNL and PL-ACMNL models with those of other extant models, on standard datasets. Naturally, PL-ACMNL has performed exceptionally well in those datasets, where both the attributes interaction and the DM's attitudinal character are crucial for the output variable, such as EMP (employee selection), LEC (lecture evaluation), DB(house evaluation in Den Bosch), etc.

7 Conclusions

A novel framework is presented to learn the decision model of a decision-maker (DM), which is a central problem to areas within computer science, operations research, marketing and econometrics. With this objective, and an emphasis on the complementarity of choice modelling and machine learning, we have shown new directions to combine these two disciplines together against the background of preference learning. The proposed approach is inspired by the human cognition process of making inferences about an individual's decision behaviour by observing his/her choices. It allows for learning a DM's decision model from observed preference information in the form of pairwise comparisons. The proposed learning approach inherits the advantages of interpretability and high prediction accuracy from the choice modelling techniques and machine learning, respectively.

The empirical study on the real world datasets show encouraging results in terms of prediction accuracy vis-a-vis state-of-the-art methods such as RankSVM, TOPSIS, POLY-kernel, and RBF-kernel. In contrast to these models, the proposed learning approach, based on the simple probability of an alternative being preferred to another, imparts robustness to the learning model. That is, the learning model is not much sensitive to the inconsistencies in a DM's own decision behaviour, or a few erroneous data. Besides, the proposed approach of *learning from preferences* is quite simple, interpretable, intuitive, and east to apply.

With regards to the future work, this work has many applications. It can be used to infer a generic model of customer choice (namely, distributions over preference lists), and subsequently to predict revenues from an offering of a particular assortment of choices. A study of this kind for a large customer base can help an organization to shape its strategy of production or marketing. The required preference data for such kind of a study can be found through CRM (customer relationship management), typically maintained by large organizations.

Learning the preference model of a DM holds a lot of significance for individuals, businesses, and governments, alike. With the knowledge of the same, businesses can bring efficiency in cross-selling strategies by focussing on the attributes of high importance; for instance by generating personalized recommendations of the potential products of a customer's interests. In the same vein, the model can be applied to analyze the citizens' preferences in response to a government regulation, say introduction of a new tax, law, vehicle emmission standards, etc.

References

[1] R. Kohavi, F. Provost, Glossary of terms, Machine Learning, 30, pp. 271–274, 1998.
[2] D. McFadden, (1973). Conditional Logit Analysis of Quantitative Choice Behavior, in Zaremmbka P. (ed.), Frontier of Econometrics, Academic Press, New York.
[3] C. Chu, (1989) A Paired Combinatorial Logit Model for Travel Demand Analysis, Proceedings of the Fifth World Conference on Transportation Research, Vol. 4, Ventura, CA, 1989, pp. 295-309.
[4] F. S. Koppelman, C-H Wen, The paired combinatorial logit model: properties, estimation and application, Transportation Research Part B: Methodological, Volume 34, Issue 2, February 2000, pp. 75-89
[5] P. Vovsha, Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area. Transportation Research Record 1607, 6–15, 1997.
[6] J. D. Lemp, K. M. Kockelman, P. Damien, The continuous cross-nested logit model: Formulation and application for departure time choice, Transportation Research Part B: Methodological, 2010, 44, 646-661.
[7] C. H. Wen, F. S. Koppelman, The generalized nested logit model. Transportation Research B 35/7, 627–641, 2001.
[8] J. Swait, Choice set generation within the generalized extreme value family of discrete choice models, Transportation Research Part B: Methodological, Volume 35, Issue 7, August 2001, pp. 643-666
[9] McFadden, D. and K. Train (2000) Mixed MNL models for discrete response, Journal of Applied Econometrics, 15, 447-470.
[10] F-M Tseng, C-Y Yu, Partitioned fuzzy integral multinomial logit model for Taiwan's internet telephony market, Omega, Volume 33, Issue 3, June 2005, pp. 267-276

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2016.2563434, IEEE Transactions on Knowledge and Data Engineering

[11] H.C.W.L. Williams, On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit, Environment and Planning, 9A, 285-344, 1977.

[12] A. Daly, and S. Zachary (1979), Improved Multiple Choice Models, in Determinants of Travel Choice, Hensher, D.A. and M. Q. Dalvi (eds.), pp. 335-357, Prager, New York.

[13] D. McFadden, Modeling the Choice of Residential Location, Transportation Research Record, No. 672, pp. 72-77, 1978.

[14] C. Daganzo, Multinomial Probit: The Theory and its Application to Demand Forecasting. Academic Press, New York, 1979.

[15] P. E. Gill, W. Murray, M. H. Wright, 1981, Practical Optimization. Academic Press: London.

[16] C. L. Hwang and K. Yoon. Multiple Attribute Decision Making: Methods and Applications. Springer, 1981.

[17] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression. In Advances in Large Margin Classifiers, Smola A, Bartlett P, Scholkopf B, Schuurmans D (eds), The MIT Press: Cambridge, Massachusetts, 2000, 115-132.

[18] Tehrani, A.F.; Weiwei Cheng; Hullermeier, E., Preference Learning Using the Choquet Integral: The Case of Multipartite Ranking, Fuzzy Systems, IEEE Transactions on , vol.20, no.6, pp.1102–1113, Dec. 2012.

[19] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), 10–18, 2009.

[21] G. Beliakov and S. James. Citation-based journal ranks: the use of fuzzy measures, Fuzzy Sets and Syst., 167(1), 101-119, 2011.

[22] H. Daniels and B. Kamp. Applications of mlp networks to bond rating and house pricing. Neural Computation and Applications, 8, 226–234, 1999.

[23] J. Furnkranz, E. Hullermeier, and S. Vanderlooy, Binary decomposition methods for multipartite ranking. In W. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, editors, Proc. of the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 359–374, Springer, 2009.

[24] C. Chen, G. Zhang, R. Tarefder, J. Ma, H. Wei, H. Guan, A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes, Accident Analysis & Prevention, Volume 80, July 2015, pp. 76-88.

[25] F. Ye, D. Lord, Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models, Analytic Methods in Accident Research, Volume 1, January 2014, pp. 72-85.

[26] R. Wang, Capacitated assortment and price optimization under the multinomial logit model, Operations Research Letters, Volume 40, Issue 6, November 2012, pp. 492-497.

[27] K. Haase, S. Muller, A comparison of linear reformulations for multinomial logit choice probabilities in facility location models, European J. of Operational Research, Volume 232, Issue 3, 1 February 2014, pp. 689-691.

[28] A. Kemal Celik, E. Oktay, A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey, Accident Analysis & Prevention, Volume 72, November 2014, pp. 66-77.

[29] T. H. Rashidi, J. Auld, A. (Kouros) Mohammadian, A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection, Transportation Research Part A: Policy and Practice, Volume 46, Issue 7, August 2012, pp. 1097-1107.

[30] S. Pulugurta, A. Arun, M. Errampalli, Use of Artificial Intelligence for Mode Choice Analysis and Comparison with Traditional Multinomial Logit Model, Procedia - Social and Behavioral Sciences, Volume 104, 2 December 2013, pp. 583-592.

[31] B. Li, The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis, Transportation Research Part B: Methodological, Volume 45, Issue 3, March 2011, pp. 461-473.

[32] C. van Campen, I. B. Woittiez, Client demands and the allocation of home care in the Netherlands. A multinomial logit model of client types, care needs and referrals, Health Policy, Volume 64, Issue 2, May 2003, pp. 229-241.

[33] A. B. Grigolon, A. W.J. Borgers, A. D.A.M. Kemperman, H. J.P. Timmermans, Vacation length choice: A dynamic mixed multinomial logit model, Tourism Management, Volume 41, April 2014, pp. 158-167.

[34] J. Maiti, A. Bhattacherjee, Evaluation of Risk of Occupational Injuries Among Underground Coal Mine Workers Through Multinomial Logit Analysis, Journal of Safety Research, Volume 30, Issue 2, Summer 1999, pp. 93-101.

[35] G. Caggiano, P. Calice, L. Leonida, Early warning systems and systemic banking crises in low income countries: A multinomial logit approach, Journal of Banking & Finance, Volume 47, October 2014, pp. 258-269.

[36] P. Davis and P. Schiraldi, The flexible coefficient multinomial logit (FC-MNL) model of demand for differentiated products, The RAND Journal of Economics, Vol. 45, 1, 2014, Pp. 32–63.

[37] U. Gazder and N. T. Ratrout, A new logit-artificial neural network ensemble for mode choice modeling: a case study for border transport Journal of Advanced Transp., 2015.

[38] N. M. Boyson, C. W. Stahel and R. M. Stulz, Hedge Fund Contagion and Liquidity Shocks, The Journal of Finance, Volume 65, Issue 5, October 2010, pp.: 1789–1816.

[39] N. Vozlyublennaia, Does Idiosyncratic Risk Matter for Individual Securities?, Financial Management, Volume 41, Issue 3, Fall 2012, pp.: 555–590

[40] P. Changpetch and D. K.J. Lin, Selection of multinomial logit models via association rules analysis Wiley Interdisciplinary Reviews: Computational Statistics, Volume 5, 1, January/February 2013, pp.: 68–77.

[41] X. Liu and Charles C. Engel, Predicting longitudinal trajectories of health probabilities with random-effects multinomial logit regression Statistics in Medicine, Volume 31, Issue 29, 20 December 2012, pp.: 4087–4101.

[42] J. Feldman and H. Topaloglu, Bounding Optimal Expected Revenues for Assortment Optimization under Mixtures of Multinomial Logits Production and Operations Management, 2015.

[43] D. A. Hensher, S. Jones and W. H. Greene, An Error Component Logit Analysis of Corporate Bankruptcy and Insolvency Risk in Australia Economic Record, Volume 83, Issue 260, March 2007, pp.: 86–103.

[44] T. Astebro and J. K. Winter, More than a Dummy: The Probability of Failure, Survival and Acquisition of Firms in Financial Distress European Management Review, Volume 9, Issue 1, Spring 2012, pp.: 1–17.

[45] Y. Bentz and D. Merunka, Neural networks and the multinomial logit for brand choice modelling: a hybrid approach Journal of Forecasting, Volume 19, Issue 3, April 2000, pp.: 177–200.

[46] R. Tchernis, M. Horvitz-Lennon and S.-L. T. Normand, On the use of discrete choice models for causal inference Statistics in Medicine, Volume 24, Issue 14, 30 July 2005, pp.: 2197–2212,

[47] S. Jackman, Bayesian Analysis of Choice Making Bayesian Analysis for the Social Sci., pp.: 379–434, 2009

[48] P. Congdon, Multinomial and Ordinal Regression Models Bayesian Statistical Modelling, Second Edition , pp.: 219–240, 2007

[49] D. A. Hensher and S. Jones, Forecasting Corporate Bankruptcy: Optimizing the Performance of the Mixed Logit Model Abacus, Vol. 43, 3, Sep. 2007, pp. 241–264,

[50] R. D. Retherford, M. K. Choe, Multinomial Logit Regression, Statistical Models for Causal Analysis, pp. 151–165, 2011

[51] P. M. Guadagni, and J. D. C. Little, A logit model of brand choice calibrated on scanner data, Marketing science, 2, 3, 1983, pp. 203-238.

[52] T. J. Gilbride, and G. M. Allenby, A choice model with conjunctive, disjunctive, and compensatory screening rules, Marketing Science, 23, 3, 2004, 391-406.

[53] W. A. Kamakura and G. Russell, A probabilistic choice model for market segmentation and elasticity structure, Journal of Marketing Research, 26, 1989, pp. 379-390.

[54] R. G. Chapman and R. Staelin, Exploiting rank ordered choice set data within the stochastic utility model, Journal of marketing research, 1982, pp. 288-301.

[55] M. Sugeno, Theory of fuzzy integral and its application. Doctorial dissertation, Tokyo Institute of Tech., 1974.

[56] G. Choquet, Theory of capacities, Ann. Inst. Fourier, Grenoble 5, pp. 131295, 1953.

[57] M. Aggarwal, Discrete Choice Models for Interactive Attributes, European J. of Oper. Res. (under review), 2015.

[58] M. Aggarwal, Attitudinal Choquet Integrals, Omega (under review), 2015.

[59] C. Lucas, T. L. Griffiths, F. Xu, C. Fawcett, A rational model of preference learning and choice prediction by children, Advances in Neural Information Processing Systems, 21, pp. 985-92, 2009.
[60] M. Peterson, 2009, An Introduction to Decision Theory. Cambridge Univ. Press: Cambridge, UK.
[61] E. Brochu, N. de Freitas, and A. Ghosh. Active preference learning with discrete choice data. In Advances in Neural Information Processing Systems 21, 2007.
[62] M. Tesch, J. Schneider, and H. Choset, Expensive Function Optimization with Stochastic Binary Outcomes. In International Conference on Machine Learning, 12831291, 2013.
[63] Y. Wang, C. Wang, W. Powell, The Knowledge Gradient with Logistic Belief Models for Binary Classification, Cornell University Library, 2015.
[64] H.J. Zimmermann, P. Zysno, Latent connectives in human decision making,Fuzzy Sets Syst. 4 (1980) 3751.
[65] M.K. Luhandjula, Compensatory operators in fuzzy linear programming withmultiple objectives, Fuzzy Sets Syst. 8 (September (3)) (1982) 245252.
[66] H. Dyckhoff, W. Pedrycz, Generalized means as a model of compensation con-nectives, Fuzzy Sets Syst. 14 (1984) 143154.

Manish Aggarwal received his B.E in CSE in 2000, M.Tech in Computer Applications and Ph.D. in Information Technology from Indian Institute of Technology (IIT) Delhi in 2006 and 2014, respectively. His research interests include multi attribute decision making, machine learning techniques, preference learning, fuzzy optimization, fuzzy decision analysis, fuzzy sets and system analysis, rough set theory, aggregation operators, non-classical logics, approximate reasoning, and plausible and analogical reasoning with applications to artificial intelligence. He has published extensively in his areas of interest in journals such as IEEE Transactions on Fuzzy Systems, Information Sciences, Applied Soft Computing, International J. of Intelligent Systems, J. of Intelligent and Fuzzy Systems, J. of Multi Criteria Decision Analysis, International J. of Machine Learning and Cybernetics, and alike.

TABLE 2: Results obtained for the different methods in terms of C-index (mean ± std. deviation (rank))

Data set	RankSVM	TOPSIS	POLY-kernel	RBF-Kernel	PL-MNL	PL-CMNL	PL-ACMNL
SJ	0.7902 ± 0.0095 (5)	0.7725 ± 0.0000 (7)	0.7855 ± 0.0111 (6)	0.8415 ± 0.0000 (4)	0.8955 ± 0.0253 (2)	0.8713 ± 0.0035 (3)	0.9214 ± 0.0008 (1)
CPU	0.9316 ± 0.0066 (6)	0.8810 ± 0.0005 (7)	0.9532 ± 0.0101 (4)	0.9480 ± 0.0080 (5)	0.9685 ± 0.0055 (2)	0.9715 ± 0.0040 (1)	0.9589 ± 0.0065 (3)
EMP	0.9245 ± 0.0098 (5)	0.9315 ± 0.0023 (4)	0.9612 ± 0.0035 (1)	0.9134 ± 0.0111 (6)	0.8514 ± 0.0036 (7)	0.9321 ± 0.0083 (3)	0.9515 ± 0.0098 (2)
MAMMO	0.8358 ± 0.0213 (4)	0.8132 ± 0.0024 (7)	0.8235 ± 0.0096 (6)	0.8311 ± 0.0081 (5)	0.8919 ± 0.0102 (1)	0.8566 ± 0.0111 (3)	0.8611 ± 0.0101 (2)
LEC	0.8714 ± 0.0111 (2)	0.7930 ± 0.0000 (7)	0.8598 ± 0.0122 (4)	0.8398 ± 0.0222 (6)	0.8699 ± 0.0091 (3)	0.8411 ± 0.0212 (5)	0.8825 ± 0.0111 (1)
CAR	0.9381 ± 0.0094 (2)	0.9211 ± 0.0005 (6)	0.9321 ± 0.0086 (3)	0.8601 ± 0.0112 (7)	0.9232 ± 0.0215 (5)	0.9466 ± 0.0101 (1)	0.9288 ± 0.0151 (4)
DB	0.9055 ± 0.0091 (5)	0.9012 ± 0.0000 (6)	0.9381 ± 0.0089 (3)	0.9285 ± 0.0255 (4)	0.8918 ± 0.0109 (7)	0.9411 ± 0.0452 (2)	0.9616 ± 0.0222 (1)
BC	0.7959 ± 0.0095 (1)	0.7152 ± 0.0011 (6)	0.7895 ± 0.0085 (2)	0.7495 ± 0.0512 (4)	0.7611 ± 0.0431 (3)	0.6825 ± 0.0581 (7)	0.7215 ± 0.0335 (5)
EMP	0.7619 ± 0.0046 (2)	0.7125 ± 0.0003 (6)	0.7765 ± 0.0113 (1)	0.6925 ± 0.0333 (7)	0.7285 ± 0.0088 (5)	0.7389 ± 0.0222 (3)	0.7311 ± 0.0242 (4)
AUTO	0.8812 ± 0.0081 (6)	0.8985 ± 0.0012 (7)	0.9132 ± 0.0056 (5)	0.9281 ± 0.0187 (3)	0.9192 ± 0.0119 (4)	0.9338 ± 0.0113 (2)	0.9415 ± 0.0315 (1)
SW	0.7255 ± 0.0092 (6)	0.7365 ± 0.0018 (4)	0.7689 ± 0.0093 (2)	0.7306 ± 0.0191 (5)	0.7215 ± 0.0129 (7)	0.7612 ± 0.0502 (3)	0.7813 ± 0.0091 (1)
CONC	0.8615 ± 0.0081 (1)	0.8125 ± 0.0062 (5)	0.8428 ± 0.0082 (2)	0.6924 ± 0.0111 (7)	0.8321 ± 0.0139 (3)	0.7614 ± 0.0110 (6)	0.8226 ± 0.0343 (4)
Avg. rank	3.75	6	3.25	5.25	4.08	3.25	2.41