A sparse nonparametric hierarchical Bayesian approach towards inductive transfer for preference modeling

Sotirios P. Chatzis *, Yiannis Demiris

Department of Electrical and Electronic Engineering, Imperial College, London, United Kingdom

A R T I C L E   I N F O

Keywords:
Preference learning
Multitask learning
Nonparametric models
Dirichlet process
Automatic relevance determination

A B S T R A C T

In this paper, we present a novel methodology for preference learning based on the concept of inductive transfer. Specifically, we introduce a nonparametric hierarchical Bayesian multitask learning approach, based on the notion that human subjects may cluster together forming groups of individuals with similar preference rationale (but not identical preferences). Our approach is facilitated by the utilization of a Dirichlet process prior, which allows for the automatic inference of the most appropriate number of subject groups (clusters), as well as the employment of the automatic relevance determination (ARD) mechanism, giving rise to a sparse nature for our model, which significantly enhances its computational efficiency. We explore the efficacy of our novel approach by applying it to both a synthetic experiment and a real-world music recommendation application. As we show, our approach offers a significant enhancement in the effectiveness of knowledge transfer in statistical preference learning applications, being capable of correctly inferring the actual number of human subject groups in a modeled dataset, and limiting knowledge transfer only to subjects belonging to the same group (wherein knowledge transferability is more likely).

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Preference learning is a research field that has attracted significant attention by the machine learning community in the last years, as it provides the effective means for modeling and predicting a human subject’s desires. Indeed, modern decision support systems (Chajewska, Koller, & Parr, 2000), recommender systems (Blei, Ng, & Jordan, 2003), and personalized service frameworks (Read, MacFarlane, & Cassey, 2002) heavily rely on sophisticated machine learning tools for preference modeling: computer games, e-commerce, marketing, and psychological support of patients are characteristic application areas that have greatly benefited from such methodologies, to name just a few (Yannakakis, Maragoudakis, & Hallam, 2009).

Existing preference learning methodologies can be generally divided into two categories: (i) methodologies based on binary preference decisions; and (ii) methodologies based on instance scoring methods (Fürnkranz & Hüllermeier, 2003). In the first case, the preference prediction problem is essentially formulated as an augmented binary classification problem, by utilizing some popular binary classification algorithm endowed with a suitable preference likelihood function (Aioli & Sperduti, 2007; Fürnkranz & Hüllermeier, 2003; Har-Peled, Roth, & Zimak, 2002; Herbrich, Graepel, Bollmann-Sdorra, & Obermayer, 1998). On the other hand, under the second approach, a utility function is introduced to directly assign ranking scores to the presented alternatives, in a regression-type fashion (Chu & Ghahramani, 2005; Crammer & Singer, 2001).

In this work, we follow the second approach: each alternative presented to a modeled subject is assigned a utility function, defined as the weighted sum of a set of basis functions, with each basis function expressing the dissimilarity of the presented alternative from a model training data point. Further, a suitable spherical Gaussian prior is imposed over the entailed weights vector of the model to conduct selection of the most informative training data points under an automatic relevance determination (ARD) scheme (Chatzis, Kosmopoulos, & Varvarigou, 2008). In other words, a sparse Bayesian inference algorithm for preference learning is eventually obtained (Bishop & Tipping, 2000). This formulation of the preference elicitation process under a sparse Bayesian learning paradigm enables our model to successfully deal with inconsistencies in the responses obtained by the modeled subjects during the model training phase, while is also very helpful in learning biases the subjects may have.

A significant issue preference learning methodologies are typically confronted with is the limited availability of labeled data suitable for model training. Typical machine learning approaches towards preference modeling consider each subject as an individual and unique case for which separate model training has to be
conducted, after acquiring some training data through interaction with that specific subject (Yannakakis et al., 2009). Indeed, obtaining appropriate model training datasets in preference learning applications requires significant time and effort from the modeled subjects. However, it is more than obvious that the average user of a real-world system would be rather reluctant to engage in a tedious training data collection procedure. Hence, in order for a preference learning methodology to cater to the actual needs of real-world systems, it must be capable of exploiting all the possibly available user information in the most efficient way, with the aim of minimizing the effort required from each individual subject for adequate model training.

To address these issues, in this work the task of learning the preferences of a modeled subject is conducted based on the assumption that each subject is not a unique and isolated case, but rather an instance or member of a group of subjects sharing a common underlying rationale in their way of making preference decisions. In other words, we seek to provide a model that can simultaneously handle the tasks of learning the preferences of multiple subjects, by regarding these tasks as a set of multiple correlated subtasks, instead of independent and isolated tasks. This way, under the proposed approach, the preference models of different subjects can regularize and influence each other, on the basis of the underlying assumption that the modeled subjects cluster into groups sharing common preference patterns.

Our approach comprises a typical example of a general machine learning framework usually referred to as multitask learning (MTL) (Caruana, 1997), or learning-to-learn (Thrun & Pratt, 1998). Multi-task learning has been the focus of much interest in the machine learning community over the last decade. The basic idea in multi-task learning is that models learned on different scenarios have parts in common (Baxter, 1995; Evgeniou, Micchelli, & Pontil, 2005; Caruana, 1997; Lawrence & Platt, 2004); in a Bayesian framework this assumption is usually formulated as a hierarchical prior shared by the models pertaining to the considered individual scenarios (Yu, Schwaighofer, Tresp, Ma, & Zhang, 2003; Yu et al., 2004; Yu, Schwaighofer, & Tresp, 2005). This hierarchical formulation of the MTL approach essentially consists of two layers: The bottom layer of the hierarchy comprises the individual, task-specific models. On the layer above, tasks are connected together via a common prior placed over the parameters of their corresponding models. Learning of the common prior is an indispensable part of the training process, and data from all tasks contribute to learning the common prior, thus enabling, in different scenarios, to have parts in common (Baxter, 1995; Evgeniou, Micchelli, & Pontil, 2005; Caruana, 1997; Lawrence & Platt, 2004); in a Bayesian framework this assumption is usually formulated as a hierarchical prior shared by the models pertaining to the considered individual scenarios (Yu, Schwaighofer, Tresp, Ma, & Zhang, 2003; Yu et al., 2004; Yu, Schwaighofer, & Tresp, 2005). This hierarchical formulation of the MTL approach essentially consists of two layers: The bottom layer of the hierarchy comprises the individual, task-specific models. On the layer above, tasks are connected together via a common prior placed over the parameters of their corresponding models. Learning of the common prior is an indispensable part of the training process, and data from all tasks contribute to learning the common prior, thus enabling, in different scenarios, to have parts in common (Baxter, 1995; Evgeniou, Micchelli, & Pontil, 2005; Caruana, 1997; Lawrence & Platt, 2004); in a Bayesian framework this assumption is usually formulated as a hierarchical prior shared by the models pertaining to the considered individual scenarios (Yu, Schwaighofer, Tresp, Ma, & Zhang, 2003; Yu et al., 2004; Yu, Schwaighofer, & Tresp, 2005).
\( G \mid \theta_0, \gamma \) \sim DP(\theta_0, \gamma) \quad (1)

\( \theta_0 \mid G \sim \text{G}, \quad n = 1 \ldots N \quad (2) \)

Integrating out \( G \), the joint distribution of the variables \( \{ \theta_n \}_{n=1}^N \) can be shown to exhibit a chaotic effect. Specifically, given the first \( N-1 \) samples of \( G \), \( \{ \theta_n \}_{n=1}^{N-1} \), it can be shown that a new sample \( \theta_N \) is either (a) drawn from the base distribution \( \theta_0 \) with probability \( \gamma / (N-1) \), or (b) is selected from the existing draws, according to a multinomial allocation, with probabilities proportional to the number of the previous draws with the same allocation (Blackwell & MacQueen, 1973). Let \( \{ \theta_n \}_{n=1}^N \) be the set of distinct values taken by the variables \( \{ \theta_n \}_{n=1}^{N-1} \). Denoting as \( f_n \) the number of values in \( \{ \theta_n \}_{n=1}^{N-1} \) that equal to \( \theta_n \), the distribution of \( \theta_N \) given \( \{ \theta_n \}_{n=1}^{N-1} \) can be shown to be of the form (Blackwell & MacQueen, 1973):

\[
p(\theta_N \mid \{ \theta_n \}_{n=1}^{N-1}, G, \gamma) = \frac{\gamma}{\gamma + N - 1} G_0 + \sum_{i=1}^{f_N-1} \frac{f_n - 1}{\gamma + N - 1} \delta_{\theta_n}
\]

where \( \delta_{\theta_n} \) denotes the distribution concentrated at a single point \( \theta_n \). These results illustrate two key properties of the DP scheme. First, the innovation parameter \( \gamma \) plays a key role in determining the number of distinct parameter values. A larger \( \gamma \) induces a higher tendency of drawing new parameters from the base distribution \( \theta_0 \); indeed, as \( \gamma \to \infty \) we get \( G \to \theta_0 \). On the contrary, as \( \gamma \to 0 \) all \( \theta_n \) tend to cluster to a single random variable. Second, the more often a parameter is shared, the more likely it will be shared in the future.

A characterization of the (unconditional) distribution of the random variable \( G \) drawn from a Dirichlet process \( DP(\theta_0, \gamma) \) is provided by the stick-breaking construction (Sethuraman 1994). Consider two infinite collections of independent random variables \( \psi \equiv \{ \psi_c \}_{c=1}^{\infty}, \{ \theta_c \}_{c=1}^{\infty} \), where the \( \psi_c \) are drawn from the Beta distribution Beta \((1, \gamma)\), and the \( \theta_c \) are independently drawn from the base distribution \( \theta_0 \). The stick-breaking representation of \( G \) is then given by Sethuraman (1994):

\[
G = \sum_{c=1}^{\infty} \pi_c(\psi) \delta_{\theta_c}
\]

where

\[
\pi_c(\psi) = \nu_c \prod_{c=1}^{\infty} (1 - \nu_c) 
\in [0, 1]
\]

and

\[
\sum_{c=1}^{\infty} \pi_c(\psi) = 1
\]

The stick-breaking representation of the DP makes clear that the random variable \( G \) drawn from a DP is discrete. It shows explicitly that the support of \( G \) consists of a countably infinite sum of atoms located at \( \theta_c \) drawn independently from \( \theta_0 \). It also apparent that the innovation parameter \( \gamma \) controls the mean value of the stick variables, \( \nu_c \), as a hyperparameter of their prior distribution; hence, it regulates the effective number of the distinct values of the drawn atoms (Sethuraman, 1994).

Under the stick-breaking representation (4) of the Dirichlet process, the atoms \( \theta_c \) drawn independently from the base distribution \( \theta_0 \) can be seen as the parameters of the component distributions of a mixture model comprising an unbounded number of component densities, with mixing proportions \( \pi_c(\psi) \). This way, DP mixture models are formulated (Antoniak, 1974). Let \( w = \{w_n\}_{n=1}^N \) be a set of variables modeled by a DPM model. Then, each one of the variables \( w_n \) is assumed to be drawn from its own probability density function \( p(w_n \mid \theta_n) \) parametrized by the parameter set \( \theta_n \). All \( \theta_n \) comply a common DP prior, and given the discreteness of \( G \), may share the same value \( \theta_n \) with probability \( \pi_c(\psi) \). Introducing the indicator variables \( z \equiv \{z_n\}_{n=1}^N \), with \( z_n = c \) denoting that \( \theta_n \) takes on the value of \( \theta_c \), the modeled set of variables \( w \) can be described as arising from the process:

\[
p(w_n \mid z_n = c, \theta_c) = p(w_n \mid \theta_c)
\]

\[
z_n \mid \pi(\psi) \sim \text{Mult}(\pi(\psi))
\]

\[
\nu_c(\gamma) \sim \text{Beta}(1, \gamma)
\]

\[
\theta_c \mid G_0 \sim G \quad (c = 1, \ldots, \infty)
\]

where \( \pi(\psi) = (\pi_c(\psi))_{c=1}^{\infty} \) is given by (5), and \( \text{Mult}(\pi(\psi)) \) is a Multinomial distribution over \( \pi(\psi) \).

### 2.2. Probabilistic preference models

Let us consider an application where a subject is presented with a number of alternatives and is asked to indicate the one that he/ she likes the most (forced-choice selection). Let, for example, the alternatives presented to the subject on the \( j \)th trial be \( \{x(j)_{c=1}^{\infty}\} \), with the \( d \)-dimensional variables \( x(j) \) taking values on \( \mathbb{R}^d \). The purpose of probabilistic preference models is to offer a way of expressing the likelihood of the subject selecting each one of the presented alternatives \( \{x(j)_{c=1}^{\infty}\} \), on the basis of the introduction of a suitable utility function, assigning a score to each one of the presented alternatives that reflects the degree of like of the subject with respect to each one of them. One of the most typical assumptions of probabilistic preference models is that the subject’s decision in such a forced-choice comparison can be modeled under the multinomial logistic (softmax) likelihood assumption (Glickman & Jensen, 2005):

\[
p(a_i = x(j)_{c=1}^{\infty}) = \frac{\exp(U(a_i \mid j))}{\sum_{c=1}^{\infty} \exp(U(a_i \mid j))}
\]

where \( U(a) \) is the utility function of the modeled subject, and \( a_i \) is the index of the alternative that the subject selected on the \( j \)th trial. Note that for pairwise comparisons, i.e., the modeled subject is always asked to select between two alternatives, the likelihood function (11) takes a form widely known as the Bradley–Terry model (Bradley & Terry, 1952).

Alternatively, the multinomial probit model may be also used as our likelihood assumption (Chu & Ghahramani, 2005). Indeed, as discussed in Rasmussen and Williams (2006), the logistic and probit models usually give similar predictions. Nevertheless, in applications where the forced-choice scenario entails more than two presented alternatives, the probit model is considerably more difficult to handle (Rasmussen & Williams, 2006). Therefore, in this work a softmax preference likelihood function will be employed.

As becomes obvious from the above description, under this probabilistic framework, learning the preference patterns of the modeled subjects essentially reduces to learning the corresponding utility functions \( U(a) \) of the modeled subjects. The contribution of this paper consists in the provision of a methodology for jointly learning the utility functions of multiple subjects, under the assumption that the modeled subjects cluster into groups of individuals sharing some common underlying rationale in their preference elicitation procedures. Additionally, we seek to provide a sparse Bayesian inference procedure for the learnt utility functions, that will allow for enhanced computational efficiency, and will endow our model with the ability to cope with the inconsistencies in the responses obtained by the modeled subjects during the model training phase.
3. Proposed approach

3.1. Model formulation

Let us consider a preference learning application, where the goal is to jointly learn the preferences of N subjects, under the MTL framework. In order to obtain some training data sets for our model, we adopt the following experimental setup: Each subject \( n = 1, \ldots, N \) engages in a set of, say, \( J \) trials. On each trial, the subject is presented with a set of alternatives, \( \{x^n(j)\}_{j=1}^{J} \), and is asked to determine the alternative he/she likes the most; let \( a^n_l \) be the index of the alternative selected by the nth subject on the jth trial. Let us also denote as \( X = \{x^n(j)\}_{n=1}^{N} \) the set of all the distinct values of the alternatives presented to the modeled subjects during the model training phase, with \( x \in \mathbb{R}^d \).

As we have already discussed, in this work we adopt a softmax preference likelihood, given by

\[
p(a^n_l = a|\{x^n(j)\}_{j=1}^{J}) = \frac{\exp[U^n(x^n(a))] }{ \sum_{a'=1}^{K} \exp[U^n(x^n(a')) ]} \tag{12}
\]

Regarding the utility function of our model, in this work we adopt a simple linear expression, given by

\[
U^n(x) = \sum_{i=1}^{J} w_{ni} \phi_i(x) \tag{13}
\]

where \( \phi_i(x) = \kappa(x, x_i) \) is a set of basis functions, expressing the dissimilarity of \( x \) from the instances in \( X \)

\[
\phi_i(x) = \kappa(x, x_i) \tag{14}
\]

and \( \kappa(\cdot, \cdot) \) is a suitable kernel function (e.g., a Gaussian RBF kernel).

As we shall explain in the following, the adoption of the utility function (13) allows for the induction of a sparse nature for our model, by application of the Bayesian automatic relevance determination methodology. ARD is a popular method for inducing sparsity in models which are linear in their parameters, by imposing a suitable prior distribution over the model parameters (Chatzis et al., 2008).

In order to facilitate the joint inference of the utility functions of the \( N \) modeled subjects under the MTL framework, while also allowing for the determination of the groups of subjects that do actually share some common patterns in their preference elicitation procedures, and, hence, information transfer between them is actually meaningful, we choose to impose the following non-parametric (DP-based) hierarchical prior over the weight vectors \( w_n \triangleq (w_{ni})_{i=1}^{J} \) of the subjects’ utility functions:

\[
p(w_n|z_n = c) = \prod_{i=1}^{J} N(w_n|0, (\beta_i^c)^{-1}) \tag{15}
\]

\[
p(\beta_i^c|\lambda, \zeta) = \mathcal{G}(\beta_i^c|\lambda, \zeta) \tag{16}
\]

\[
z_i|\pi(\pi) \sim \text{Mult}(\pi(\pi)) \tag{17}
\]

\[
\pi_c(\pi) = \kappa_c \prod_{c=1}^{C-1} (1 - \nu_c ) \tag{18}
\]

\[
\nu_c/\gamma \sim \text{Beta}(1, \gamma) \tag{19}
\]

\[
\gamma|\eta_1, \eta_2 \sim \text{G}(\eta_1, \eta_2) \tag{20}
\]

with \( C = 1, \ldots, \infty \).

Note that the conjugate prior (15) imposed on the weight vectors \( w_n \) conditional on the cluster their corresponding subjects belong to has the appropriate form that allows for the conduction of the ARD mechanism in the context of our model, thus giving rise to a sparse Bayesian inference algorithm (Chatzis et al., 2008). The notion of ARD is to continually create new components while detecting when a component model starts to overfit. The overfit manifests itself as a precision hyperparameter posterior mean tending to infinity, indicating that only a single data value is being modeled by the component. Hence, in the case of our model, the ARD mechanism is implemented by imposing a hierarchical prior over the weight vectors \( w_n \), with the width of each prior being controlled by a Gamma-distributed precision hyperparameter, \( \beta_i^c \), as illustrated in Eqs. (15) and (16). If one of these precisions tends to infinity, \( \beta_i^c \rightarrow \infty \), then the outgoing weights \( w_{ni} \) for the nth subject, given that it belongs to the cth cluster of the modeled subjects, will have to be very close to zero in order to maintain a high likelihood under this prior, which in turn leads the subject’s utility function to ignore the contribution of the corresponding term \( w_{ni} \phi_i(x) \). Therefore, the corresponding training data point \( x_i \) is effectively “switched off” w.r.t. the model inference procedure for a modeled subject attributed to the cth subjects cluster, thus inducing a sparse nature for our model.

3.2. Model inference

Inference for hierarchical Bayesian models is typically conducted either by means of variational Bayes (e.g., Blei & Jordan, 2006), or Monte Carlo techniques (e.g., Neal, 2000). Here, we prefer a variational Bayesian approach, due to its considerably better scalability in terms of computational costs, which becomes of a major importance when considering real-life applications. Bayesian inference involves introduction of a set of appropriate priors over the model parameters, and derivation of the corresponding (approximate) posterior densities.

Our variational Bayesian inference formalism for the proposed model consists in derivation of a family of variational posterior distributions \( q(\cdot) \) which approximate the true posterior distribution over the infinite sets \( \pi = (\nu_1, \ldots, \nu_C) \) and \( \beta_i^c | i = 1, \ldots, J \) the innovation parameter \( \gamma \), and the weights \( \{w_{ni}\}_{i=1}^{J} \) of the subjects’ utility functions. Apparently, under this infinite-dimensional setting, Bayesian inference is not tractable. For this reason, we employ a common strategy in the literature of nonparametric Bayesian models, formulated on the basis of a truncated stick-breaking representation of the DP (Blei & Jordan, 2006). That is, we fix a value \( K \) and we let the variational posterior over the \( \nu_i \) have the property \( \nu_{i+1} = 1 \). In other words, we set \( \pi_c(\pi) \) equal to zero for \( c > K \). Note that, under this setting, the treated nonparametric model involves a full DP prior; truncation is not imposed on the model itself, but only on the variational distribution to allow for a tractable inference procedure. Hence, the truncation level \( K \) is a variational parameter which can be freely set, and not part of the prior model specification.

Let \( \Psi = \{ \pi, \gamma, \{z_n\}_{n=1}^{N}, \{\beta_i^c\}_{i=1}^{J}, \{w_{ni}\}_{i=1}^{J} \} \) be the set of all the hidden variables and unknown parameters of the proposed hierarchical model over which a prior distribution has been imposed, and \( H \) be the set of the hyperparameters of the imposed priors, \( H = (\lambda, \zeta, \eta_1, \eta_2) \). Variational Bayesian inference consists in the introduction of an arbitrary distribution \( q(\Psi) \) to approximate the actual posterior \( p(\Psi|X;H) \) which is computationally intractable (Bishop, 2006). Under this scheme, the log marginal likelihood (log evidence) of the model with respect to a given training dataset \( X \), yields (Jordan, Ghahramani, Jaakkola, & Saul, 1998)

\[
\log p(X) = \mathcal{L}(q) + \text{KL}(q||p) \tag{21}
\]

where the quantity \( \mathcal{L}(q) \), usually referred to as the variational free energy, is defined as

\[
\mathcal{L}(q) = \int d\Psi q(\Psi) \log \frac{p(X, \Psi|H)}{q(\Psi)} \tag{22}
\]

and \( \text{KL}(q||p) \) stands for the Kullback–Leibler (KL) divergence between the (approximate) variational posterior, \( q(\Psi) \), and the actual posterior, \( p(\Psi|X;H) \) of the model.
KL(q∥p) = −∫dΨq(Ψ) log \frac{p(Ψ|X, H)}{q(Ψ)} \ (23)

Since KL divergence is nonnegative, L(q) forms a strict lower bound of the log evidence, and would become exact if q(Ψ) = p(Ψ|X, H). Hence, by maximizing this lower bound L(q) so that it becomes as tight as possible, not only do we minimize the KL-divergence between the true and the variational posterior, but we also implicitly integrate out the unknowns Ψ.

For computational convenience, we assume that the variational posterior q(Ψ) factors similar to the prior p(Ψ), yielding

q(Ψ) = q(γ) \prod_{i=1}^{N} q(w_{ni}) \prod_{c=1}^{K} q(v_{ic}) \prod_{i=1}^{l} q(\beta_{i}) \ (24)

This factorization assumption for the model posterior is a common approach in variational Bayesian inference techniques and is often referred to as the mean-field approximation (Jordan et al., 1998). Then, based on the definition (22), and ignoring the entailed constant terms, the variational free energy of the postulated model takes the form

L(q) = \sum_{i=1}^{N} \sum_{c=1}^{K} \int d\beta_{i} q(\beta_{i}) \log \frac{p(\beta_{i}|\lambda_{i}, \zeta)}{q(\beta_{i})} + \int d\gamma q(\gamma) \{ \log \frac{p(\gamma|\eta_{1}, \eta_{2})}{q(\gamma)} \} + \sum_{c=1}^{K} \int d\nu_{c} q(\nu_{c}) \log \frac{p(\nu_{c}|\gamma)}{q(\nu_{c})} \} + \sum_{c=1}^{K} \sum_{i=1}^{N} q(z_{i}=c) \left\{ \int d\nu_{c} q(\nu_{c}) \log p(z_{i}=c|\nu_{c}) - \log q(z_{i}=c) \right\} + \sum_{i=1}^{N} \sum_{k=1}^{l} \int dw_{ni} q(w_{ni}) \left\{ \int d\beta_{i} q(\beta_{i}) \log \mathcal{N}(w_{ni}|0, (\beta_{i})^{-1}) \right\} + \sum_{i=1}^{N} \sum_{k=1}^{l} \int dw_{ni} q(w_{ni}) \log p(w_{ni}|x_{ni}(j)^{k}) \} - \sum_{i=1}^{N} \int dw_{ni} q(w_{ni}) \log p(w_{ni}) \ (25)

Derivation of the variational posterior distribution q(Ψ) involves maximization of the variational free energy L(q) over each one of the factors of q(Ψ) in turn, holding the others fixed, in an iterative manner (Chandler, 1987). By construction, this iterative, consecutive updating of the variational posterior distribution is guaranteed to monotonically and maximally increase the free energy L(q), which functions as the convergence criterion of the derived inference algorithm for the postulated model (Chatzis et al., 2008). Let us denote as (.) the posterior expectation of a quantity. Taking the derivatives of L(q) w.r.t. each one of the factors of q(Ψ), as determined in (24), we yield the following variational posteriors:

3.2.1. q(\nu_{c})

q(\nu_{c}) = Beta(\alpha_{c,1}, \alpha_{c,2}) \ (26)

where

\alpha_{c,1} = 1 + \sum_{i=1}^{N} q(z_{i}=c) \ (27)

and

\alpha_{c,2} = \gamma + \sum_{c'=c+1}^{K} \sum_{i=1}^{N} q(z_{i}=c') \ (28)

3.2.2. q(\gamma)

q(\gamma) = G(\gamma|\tilde{\eta}_{1}, \tilde{\eta}_{2}) \ (29)

where

\tilde{\eta}_{1} = \eta_{1} + K - 1 \ (30)

and

\tilde{\eta}_{2} = \eta_{2} - \sum_{c=1}^{K-1} \left[ \psi(\alpha_{c,2}) - \psi(\alpha_{c,1} + \alpha_{c,2}) \right] \ (31)

with \psi(\cdot) denoting the digamma function.

3.2.3. q(\beta_{i})

q(\beta_{i}) = \mathcal{G}(\beta_{i}|\tilde{\xi}_{i}, \tilde{\zeta}_{i}) \ (32)

where

\tilde{\xi}_{i} = \lambda + \frac{1}{2} \sum_{i=1}^{N} q(z_{i}=c) \ (33)

and

\tilde{\zeta}_{i} = \xi + \frac{1}{2} \sum_{i=1}^{N} q(z_{i}=c) \langle w_{ni} \rangle \ (34)

3.2.4. q(w_{ni})

From (25), the equation \frac{\partial L(q)}{\partial q(w_{ni})} = 0 yields

\log q(w_{ni}) \propto \sum_{i=1}^{N} \sum_{c=1}^{K} q(z_{i}=c) \left\{ \frac{1}{2} w_{ni}(\beta_{c}) + \frac{1}{2} \log (\beta_{c}) \right\} + \sum_{j=1}^{l} \log p(w_{ni}|x_{ni}(j)^{k}) \ (35)

Apparently, the variational posterior given by (35) is not of Gaussian form, since the preference likelihood p(w_{ni}|x_{ni}(j)^{k}) is not Gaussian. To address this issue, and allow for the derivation of a variational posterior over q(w_{ni}) taking a more convenient form, we resort to the popular Laplace approximation (Bishop, 2006): Taking these second-order Taylor expansion of \log q(w_{ni}) around its mode, we yield

q(w_{ni}) \approx \mathcal{N}(w_{ni}|\mu_{ni}, \Sigma_{ni}) \ (36)

where

\mu_{ni} = \text{argmax}_{w_{ni}} \{ \log q(w_{ni}) \} \ (37)

and

\Sigma_{ni}^{-1} = - \nabla \nabla \log q(w_{ni}) \ |_{w_{ni} = \mu_{ni}} \ (38)

In this work, computation of the posterior mode \mu_{ni} is conducted by application of the Newton–Raphson algorithm, initialized at \mu_{ni} = 0 and with iteration

\mu_{ni}^{(l+1)} = \mu_{ni}^{(l)} - \left( \nabla \nabla \log q(w_{ni}) \bigg|_{w_{ni} = \mu_{ni}^{(l)}} \right)^{-1} \nabla \log q(w_{ni}) \bigg|_{w_{ni} = \mu_{ni}^{(l)}} \ (39)

This requires availability of the expressions of \nabla \nabla \log q(w_{ni}) and \nabla \nabla \log q(w_{ni}), which can be found in the Appendix A.

Table 1

<table>
<thead>
<tr>
<th>Ablalone dataset: results obtained from models trained using data from 10 trials per subject.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First scenario</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Error rate (%)</td>
</tr>
<tr>
<td># Clusters</td>
</tr>
<tr>
<td>Clustering accuracy (%)</td>
</tr>
<tr>
<td>Second scenario</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Error rate (%)</td>
</tr>
<tr>
<td># Clusters</td>
</tr>
<tr>
<td>Clustering accuracy (%)</td>
</tr>
</tbody>
</table>
Finally, regarding the posterior probabilities of cluster assignment of the modeled subjects, we have

\[ q(z_n = c) / \sim p_c(v) / \sim p(w_n | z_n = c) \leq 40 \]

where

\[ \tilde{p}_c(v) \triangleq \exp (\log p_c(v)) = \exp \left[ \sum_{c=1}^{C-1} \log (1 - \nu_c) + \log \nu_c \right] \]

and

3.2.5. \( q(z_n = c) \)

Finally, regarding the posterior probabilities of cluster assignment of the modeled subjects, we have

\[ q(z_n = c) \propto \tilde{p}_c(v) p(w_n | z_n = c) \]

where

\[ \tilde{p}_c(v) \triangleq \exp (\log p_c(v)) = \exp \left[ \sum_{c=1}^{C-1} \log (1 - \nu_c) + \log \nu_c \right] \]

and

Fig. 1. Abalone dataset: evolution of the error rates obtained by the evaluated methods as a function of the number of training examples: (a) thee-alternatives scenario; (b) four-alternatives scenario.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>First scenario</th>
<th>Second scenario</th>
<th>Birlutiu et al. (2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate (%)</td>
<td>35.90</td>
<td>45.45</td>
<td>38.49</td>
</tr>
<tr>
<td># Clusters</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Clustering accuracy (%)</td>
<td>80.0</td>
<td>80.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>

Housing dataset: results obtained from models trained using data from 10 trials per subject.
The data points are presented to each subject in a random order. We consider two groups of human subjects: the first group consists of five human subjects, who always select the data point \( x \) yielding the highest target value among the presented alternatives; the second group consists of the human subjects who always select the data point \( x \) yielding the lowest target value among the presented alternatives. In total, ten different randomly selected triplets of data points are presented to each one of the ten human subjects, and the so-obtained data is used for model inference. Subsequently, the model is evaluated using another 2000 randomly selected triplets as our test data.

2. The second scenario differs from the first one in the sense that, in each trial, each human subject is presented with four alternatives instead of only three.

The variational truncation threshold of the Dirichlet process is taken equal to \( K = 10 \) in each case, that is equal to the number of the modeled subjects (which, indeed, is the maximum possible number of subject groups). Apart from our method, we also evaluate the Gaussian process-based preference learning method of Birlutiu, Groot, and Heskes (2010) on the same scenarios. Both algorithms are evaluated using Gaussian RBF kernels. In Table 1, we present the performance obtained by the evaluated methods in terms of the yielded average preference prediction error rates over the modeled human subjects. As we observe, the proposed method yields a significantly improved preference prediction error rate compared to the considered alternative approach.

To examine where this clear advantage of the proposed algorithm actually stems from, in Table 1 we also provide the number of clusters obtained by the proposed approach. As we observe, our approach estimates the actual number of subject clusters (groups) with complete success in both scenarios; as we also notice, the subject clustering result of our method turns out to be extremely accurate, since only one subject is misclassified in the first scenario, whereas all subjects are correctly classified in the second scenario.

This finding seems to provide a rather deep insight into the reasons why our method performs much better than the competition: under the assumption that all the modeled human subjects share a common prior (rationale) over their preference elicitation processes, the postulated model combines the information collected from the different modeled human subjects in a rather misleading way; hence, the obtained prediction performance of the model turns out to be rather unsatisfactory. On the contrary, considering that different subjects may belong to different, a priori unknown groups of subjects sharing some common rationale in their preference elicitation procedures, the postulated model eventually limits information transfer only within the inferred groups of modeled subjects, thus maximizing the potential gain from information exchange by determining which bits of the available information are actually transferable between each pair of modeled human subjects.

Finally, to investigate how the performance of the evaluated methods would be affected by an increase in the number of the available training data points, and, hence, assess in a quantitative manner the actual contribution of inter-subject information transfer in the effectiveness of the preference learning algorithm, we repeat the previous experiments considering different numbers of training examples (sets of three or four alternatives presented to the modeled subjects to choose from, respectively). In Figs. 1(a) and (b), we illustrate how the obtained average error rate of the evaluated algorithms changes with the number of training triplets or quadlets per subject, respectively.

As we observe, in cases of ample availability of training examples, both algorithms seem to perform equally well; this probably indicates that, in such cases, information transfer between the
modeled human subjects does not bring about a significant benefit for the performance of the preference learning algorithms, nor does the proper determination of the specific groups of human subjects between which information is actually transferable. Additionally, as the number of training examples decreases, we observe that the proposed algorithm continues to perform comparably well, contrary to the considered alternative method which undergoes a notable performance deterioration. This finding implies that, while information transfer between the modeled human subjects becomes of vital importance when training data availability is limited, it is even more significant, in order for the information transfer to be of any advantage for the preference learning algorithm, to obtain a measure of information transferability between the modeled human subjects, and limit information transfer only inside groups of subjects within which information is likely to be transferable.

4.1.2. Housing, Machine CPU, and Pyrimidines

Further, we repeat the previous experiments using the rest of the considered datasets from the UCI machine learning repository, that is Housing, Machine CPU, and Pyrimidines. The obtained results are illustrated in Tables 2–4, and Figs. 2–4, respectively. Once again, we observe that trying to infer the groups of subjects...
exhibiting similar preference patterns, and limiting information transfer within such groups, turns out to be of significant benefit to the preference learning algorithm when training data availability is clearly limited. It is also characteristic that, as the number of the available training data points increases, the significance of information transfer between subjects wanes, and the same holds for the competitive advantages stemming from the limitation of information transfer within subject groups exhibiting similar patterns in their preference elicitation procedures.

4.2. Application to music recommendation

In this experiment, we used a music database consisting of 1000 music pieces. The 85% of our database comprised Western classical music genres: opera, orchestral music, and chamber music. Additionally, a small set of pop music pieces (comprising the rest 15% of our database) was also included. To represent the obtained music signals, we used the mono channel of the raw music signals to compute the following features:
12 mel-frequency cepstral coefficients (MFCC), and their first derivatives, the means and variances of the signal zero crossing rates, the spectral center of gravity, the spectral rolloff; and the spectral flux of the music signals.

This way, a 32-dimensional feature vector was eventually obtained from each music piece.

In our experiments, we considered a set of 30 different human subjects. Each one was presented with 6 options (music pieces randomly selected from the database, including at least one piece...
of pop music and one piece of classical music in each case), and were asked to indicate the one they liked the most. This experiment was repeated 100 times for each human subject: the first 50 datasets were used for model training, and the rest for testing. The modeled subjects can be divided into two categories: 10 of them have a clear preference for classical music; the rest clearly prefer pop music.

The obtained results are the following: the trained model determines three groups of human subjects; two of them pertain to subjects that prefer classical music, whereas the third one includes the subjects that prefer pop music. Considering the two clusters pertaining to classical music fans as one group, and the third cluster as another distinct group, the subject classification result of our model turns out to be 100% accurate. The obtained preference prediction error rates of the evaluated methods are depicted in Table 5. Both algorithms were evaluated using a Gaussian RBF kernel with unitary variance. As we observe, limiting information transfer only within groups of subjects the preference patterns of which seem to cluster together, we obtain an error reduction of as much as 42.62%. Therefore, we once again observe the clear advantages that stem from limiting information transfer only within groups of subjects exhibiting some common underlying rationale in their preference elicitation procedures.

5. Conclusions

In this work, we introduced a novel sparse nonparametric hierarchical Bayesian approach towards inductive transfer for preference modeling. Our approach is based on the consideration of a hierarchical Bayesian formulation of the multitask learning framework, with the common prior shared among the tasks taken as a Dirichlet process mixture model with a special spherical Gaussian form for its component densities, which allows for the induction of a sparse nature for our model, by application of the ARD mechanism. This way, the proposed model conducts the data-driven discovery of groups of human subjects exhibiting some common underlying rationale in their preference elicitation procedures, and allows for information sharing among the human subjects of each group in the context of their model inference procedures. The efficacy of our novel approach was experimentally validated by considering both experiments with a synthetic dataset, and a real-world application dealing with automatic music recommendation.

Acknowledgments

The authors thank Dr. D. Kosmopoulos for his valuable contribution to the conduction of the music recommendation experiments. This work has been funded by the EU FP7 ALIZ-E project (grant 248116).

Appendix A

Regarding the posterior expected values in (26)–(41), we have

\[ \langle \gamma \rangle = \frac{\bar{n}_1}{\bar{n}_2} \tag{44} \]

\[ \langle \log \psi \rangle = -\langle \log \psi \rangle - \langle \log \psi \rangle - \langle \log \psi \rangle \tag{45} \]

\[ \langle \beta \rangle = \frac{\bar{w}_c}{\bar{w}_c} \tag{46} \]

\[ \langle \log \beta \rangle = \langle \log \beta \rangle \tag{47} \]

and

\[ \langle \psi \rangle = \langle \log \psi \rangle + \langle \log \psi \rangle \tag{48} \]

Concerning the expression of the gradient \( \nabla \log p(\mathbf{w}_n) \), we have

\[ \nabla \log p(\mathbf{w}_n) = -\sum_{c=1}^{C} \langle \beta \rangle \text{diag} \left[ (\beta / \langle \beta \rangle)_{c=1} \right] \mathbf{w}_{c} + \sum_{c=1}^{C} \nabla \log p\left( \theta_{c}^{(c)} \right) \left( \beta / \langle \beta \rangle \right)_{c=1} \tag{50} \]

where

\[ \nabla \log p\left( \theta_{c}^{(c)} \right) \left( \beta / \langle \beta \rangle \right)_{c=1} = \left( \text{grad}_{1} + \text{grad}_{2} \right) \times \exp \left[ -\mathbf{U}(\theta_{c}^{(c)}) \right] \sum_{c=1}^{C} \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{51} \]

with

\[ \text{grad}_{1} = \left( \phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \right) \sum_{c=1}^{C} \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{52} \]

\[ \text{grad}_{2} = -\exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \sum_{c=1}^{C} \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{53} \]

and \( \phi(\theta_{c}^{(c)}) \).

Finally, the expression of the Hessian \( \nabla \nabla \log p(\mathbf{w}_n) \) is given by

\[ \nabla \nabla \log p(\mathbf{w}_n) = \sum_{c=1}^{C} \langle \beta \rangle \text{diag} \left[ (\beta / \langle \beta \rangle)_{c=1} \right] \mathbf{w}_{c} + \sum_{c=1}^{C} \nabla \nabla \log p\left( \theta_{c}^{(c)} \right) \left( \beta / \langle \beta \rangle \right)_{c=1} \tag{54} \]

where

\[ \nabla \nabla \log p\left( \theta_{c}^{(c)} \right) \left( \beta / \langle \beta \rangle \right)_{c=1} = \left( \text{hessian}_{1} + \text{hessian}_{2} \right) \times \exp \left[ -\mathbf{U}(\theta_{c}^{(c)}) \right] \sum_{c=1}^{C} \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{55} \]

with

\[ \text{hessian}_{1} = \frac{\phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right]}{\sum_{c=1}^{C} \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right]} \tag{56} \]

\[ \text{hessian}_{2} = -\exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \sum_{c=1}^{C} \phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{57} \]

\[ \text{hessian}_{3} = \frac{\phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \phi(\theta_{c}^{(c)})}{\sum_{c=1}^{C} \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right]} \tag{58} \]

\[ \text{hessian}_{4} = -2 \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \sum_{c=1}^{C} \phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \phi(\theta_{c}^{(c)}) \tag{59} \]

\[ \text{hessian}_{5} = -\sum_{c=1}^{C} \phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{60} \]

\[ \text{hessian}_{6} = 2 \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \sum_{c=1}^{C} \phi(\theta_{c}^{(c)}) \exp \left[ \mathbf{U}(\theta_{c}^{(c)}) \right] \tag{61} \]
References