Abstract

1. What is the bigger picture? 2. Dissertation purpose 3. Research method 4.

Key results 5. Practical Implications

Typically 3 paragraphs: 1. the introduction of the challenge problem, 2. thesis, 3. contribution

Abstract

En puhu suomea!

Preface

TO fill.

Espoo, April 11, 2019,

Pedram Daee

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Pedram Daee, Joel Pyykkö, Dorota Głowacka, and Samuel Kaski. Interactive Intent Modeling from Multiple Feedback Domains. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, Sonoma, California, USA, 71–75, March 2016.

II Pedram Daee*, Tomi Peltola*, Marta Soare*, and Samuel Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning*, 106, 9-10, 1599–1620, 2017.

III Iiris Sundin*, Tomi Peltola*, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daee, Chen He, Baris Serim, Aki Havulinna, Caroline Heckman, Giulio Jacucci, Pekka Marttinen, and Samuel Kaski. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34, 13, i395–i403, 2018.

IV Pedram Daee*, Tomi Peltola*, Aki Vehtari, and Samuel Kaski. User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, Tokyo, Japan, 305–310, March 2018.

V Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, Benjamin Blankertz. Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval. *Journal of the Association for Information Science and Technology*, 2019.

Author's Contribution

Publication I: "Interactive Intent Modeling from Multiple Feedback Domains"

The author had the main responsiblity in problem formulation and modeling. The author designed and implemented the simulation experiment. Joel Pyykkö and the author built the system for user studies and conducted them together. The author wrote the initial draft of the manuscript, after which all co-authors joined for revisions.

Publication II: "Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction"

The ideas and experiments in this article were designed jointly (the first three authors contributed equally). The author had the main responsibility in the derivation of the sequential experimental design and implementation of the experiments. Dr. Tomi Peltola derived and implemented the posterior approximation. The manuscript was written jointly.

Publication III: "Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge"

The author contributed on formulation of the sequential experimental design and implementation of a portion of the early version of the experiments. The author made comments to the manuscript in preparation.

Publication IV: "User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction"

The ideas and experiments in this article were designed jointly (the first two authors contributed equally). The author designed and implemented the user study. Dr. Tomi Peltola had the main responsibility of the model formulation. The first two authors wrote the initial draft of the manuscript, after which all co-authors joined for revisions.

Publication V: "Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval"

The author had the main responsibility in design and implementation of the interactive intent modelling and information retrieval system, and writing of the correspnding sections. All the authors contributed to paper revisions.

1. Introduction

- We can keep this empty and start from next. Alternatively we can write some stuff here and have only one section as "Contributions and organization of the thesis". See which one works better for you

Description of the Problem. Whether it is an everyday user searching for an application in her mobile phone or a doctor working with a cancer diagnostic system, humans and machines are increasingly interacting with each other. The goal of my thesis is to improve this interaction by incorporating a model of the human user in the system they are interacting with. In particular, I consider the family of problems where the human and machine interact to solve a prediction problem. Such problems can include personalized search activity of a user or medical prediction about the response of a cancer drug. An important common factor in both these scenarios is that the number of labeled data (training data) that the machine can use to make predictions, is usually very few compared to the dimension of search space. This results in ill-posed statistical learning since there are limits in how low in sample size statistical methods can go.

Objective. The objective of the thesis is to tackle the limited data problem by directly integrating a model of the human user into the modelling loop. The user can then provide more information about the problem (for example by giving feedback in response to the machine's query) that can help to improve the prediction performance. For example, in precision medicine obtaining additional data can be extremely costly or even impossible but the human knowledge, e.g., practitioners feedback, can be available and used to improve the prediction. My thesis proposes new probabilistic machine learning models to learn the human intent in interaction, new user models to learn about the knowledge of the human and accounting for common human biases, and new prediction models to employ and extract human knowledge to improve a task.

Methodology. I design probabilistic machine learning models to tie the data model (prediction model from the training data) to the user model (the model of the human) in a unified way. After doing so, we use Bayesian inference to find

the posterior, i.e, distribution of the unknown parameters related to data and user, giving the training data and user interaction. The posterior inference, in most cases, does not have an analytical solution. We use different approximation methods, such as expectation propagation and variational inference, to handle the computation. A core characteristic of the formulation is that the model adapts to the feedback obtained from the user and it sequentially integrates every piece of information before deciding on the next query for the user. This is important as humans can only answer limited number of queries (out of several thousands). The query selection is naturally formulated as experimental design problem, aiming at maximizing the most information gained, or multi-armed bandit problem, aiming at finding the most important piece of information, depending on the targeted task. For evaluation and testing, we conduct carefully designed user studies with real human.

1.1 Motivation

- Thinking back, I think my main motivation should be on how to tackled limited interaction problem. In iui2016 we did it by adding new likelihoods, in JASIS by implicit interaction, in ML and IUI 18 by assuming strong priors.

- A) Small number of data ->add users

- B) user is the source of data

- Limited interaction: 1) need to be smart about what question to ask from the user 2) Should be able to consider all potential sources of feedback on different components of the interaction framework.

1.2 Research questions and contributions

This thesis investigates methods to tackle the limited interaction challenge in interactive machine learning scenarios where there is few labeled data available compared to the dimension of the problem, or a human user is provider of the labeled data.

RQ1 – *Can we exploit new sources of interaction as new learning signals from the human user to improve interactive intent modelling?*

Publications I and V contribute to this research question by proposing models to incorporate new types of user feedback to amend the limited feedback in exploratory information seeking tasks. The information seeking tasks considered are document search scenarios where a user needs to sequentially provide

relevance feedback to suggested keywords in order to find the targeted document. The problem is usually modelled as a multi-armed bandit problem with the goal of finding the most relevant document with minimum interaction. In particular, Publication I proposes methods to additionally account for user relevance feedback on documents and couples relevance feedback on both keywords and documents by assuming a shared underlying latent model connected through a probabilistic model of the relationship between keywords and documents. Thompson sampling on the posterior of the latent intent was then used to recommend new documents and keywords in the next iteration. Publication V investigates the use of implicit relevance feedback from neurophysiology signals for effortless information seeking. The work contributes by demonstrating how to integrate this inherently noisy implicit relevance feedback combined with scarce explicit feedback. In particular, a parameter for controlling the accuracy of the feedback was introduced in the likelihood function which would have different prior based on the nature of the feedback (implicit or explicit). Similar to Publication I, Thompson sampling was used to control the exploration and exploitation balance of the recommendations. Both publications were evaluated by user studies in realistic information seeking tasks.

RQ2 – *Can human knowledge about the underlying data model be elicited to improve the prediction performance in "small n large p" tasks?*

Publications II and III contribute to this research question. Publication II proposes a framework for user knowledge elicitation as a probabilistic inference process, where the user knowledge is sequentially queried to improve predictions. In particular, sparse linear regression is considered as the data model with access to only few high-dimensional training data. It is assumed that there are human experts who have knowledge about the relevance of the covariates, or of values of the regression coefficients and can provide this information to the data model if queried. The work contributes by an algorithm and computational approximation for fast and efficient interaction, which sequentially identifies the most informative queries to ask from the user. Publication III, builds on Publication II by adding user knowledge about direction of relevance of covariates and applying the method in important applications of precision medicine with the goals of predicting the efficacies of different treatments using high-dimensional genomic measurements. Both publications were evaluated by extensive simulations and user studies. Source codes for methods presented in Publications II and III, and user study data from Publication II are available at `https://github.com/HIIT/knowledge-elicitation-for-linear-regression` and `https://github.com/AaltoPML/knowledge-elicitation-for-precision-medicine`.

RQ3 – *Is it enough to incorporate human knowledge directly in the model as explained in RQ2, or could it be beneficial to account for underlying cognitive procedures [TODO: rational knowledge updates?] that humans undergo during the interaction?*

Publication IV contributes to this research question by modelling the knowledge provider, here the human user, as a rational agent that updates its knowl-

edge about the underlying prediction task during the interaction. The design of the system is then critical, since the elicited user knowledge cannot be assumed to be independent from the data model knowledge coming from the training data. If not accounted properly, knowledge elicitation can then lead to double use of data and overfitting, if the user reinforces noisy patterns in the data. We propose a user modelling methodology, by assuming simple rational behaviour, to correct the problem and evaluate the method in a user study. Source code and user study data are available at `https://github.com/HIIT/human-overfitting-in-IML`.

1.3 Organization of the thesis

The organization of the thesis is as follows. Chapter 2 provides an overview of probabilistic modelling and introduces our approach of modelling the user and data as a joint probabilistic model. Chapter three investigates the purpose of interactions and utility functions that need to be used for efficient interaction to minimize the effort of the human user. The fourth chapter summarizes Publications I-IV. Chapter five concludes the thesis and provides discussions for future works.

2. Probabilistic modelling

- Explain general Bayesian modeling (posterior, likelihood, prior). represent uncertainty using probability distribution and then using Bayesian theorem for the inference, finding the distribution of the parameters of interest (unknown parameters), giving the observed variables and model assumptions.

- Our approach, throughout this thesis is to model both data and user as probabilistic models (figure XXX).... The role of model (prior) assumptions is crucial then since in all scenarios considered in this thesis there are only few observed data available which is usually smaller than the dimension of the problem (e.g., feature dimension)...

- Idea: Have a picture per section where there is a cloud on top of user and system head and in it there is the plate diagram of the particular model. Then talk about the interaction. In the first introduction section have the same figure but just write down (or explain) user probabilistic model and data probabilistic model (priors) and arrows for interactions? and then refer to later chapters for the type of interaction

- Then explain the two possible likelihoods and show a diagram where there is a section for data modeling and there is a section for user modeling that are connected through a shared latent parameter and then each have their own models. Then explain that we will go through details of the data and user model in the next two subsections. My contribution comes from adding data likelihoods and user priors and having interaction.

2.1 Modelling the data

- Defining appropriate likelihood and defining appropriate priors

2.1.1 Bayesian Linear regression

2.2 Modelling the user

- Our general approach is to consider new likelihood function for the user feedback and put appropriate priors on top of that. Explain the feedback likelihoods?

3. User interaction with the probabilistic model

- Here have the same figure, explain how user feedback is gathered. explain that we consider two sets of works. One where there is external data other than user feedback and one where the user feedback IS the data. Then explain our interaction goals and acquisition functions and experimental design.

4. Applications and results OR Summary of the Publications

- Go through the papers and tasks.

5. Discussion

- Motivate future works such as next level of user modelling. Similar to overfitting work, but assuming a more active version of the user -> ATOM

References

[1] Donald E. Knuth. *The TEXbook*. Addison-Wesley, 1984.

[2] WikiBooks. LaTeX. http://en.wikibooks.org/wiki/LaTeX/, 2008.

Publication I

Pedram Daee, Joel Pyykkö, Dorota Głowacka, and Samuel Kaski. Interactive Intent Modeling from Multiple Feedback Domains. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, Sonoma, California, USA, 71–75, March 2016.

Interactive Intent Modeling from Multiple Feedback Domains

Pedram Daee[1], Joel Pyykkö[2], Dorota Głowacka[2], and Samuel Kaski[1]
Helsinki Institute for Information Technology HIIT
[1]Aalto University, Department of Computer Science
[2]University of Helsinki, Department of Computer Science
firstname.lastname@hiit.fi

ABSTRACT
In exploratory search, the user starts with an uncertain information need and provides relevance feedback to the system's suggestions to direct the search. The search system learns the user intent based on this feedback and employs it to recommend novel results. However, the amount of user feedback is very limited compared to the size of the information space to be explored. To tackle this problem, we take into account user feedback on both the retrieved items (documents) and their features (keywords). In order to combine feedback from multiple domains, we introduce a *coupled multi-armed bandits* algorithm, which employs a probabilistic model of the relationship between the domains. Simulation results show that with multi-domain feedback, the search system can find the relevant items in fewer iterations than with only one domain. A preliminary user study indicates improvement in user satisfaction and quality of retrieved information.

Author Keywords
Exploratory search; intent modeling; multi-armed bandits; relevance feedback; probabilistic user models.

ACM Classification Keywords
H.3.3 Information Search and Retrieval: Relevance feedback; I.2.6 Learning.

INTRODUCTION
The dominant information retrieval paradigm relies on the user's ability to form a precise query, which is difficult at least in the about 50% of search sessions where the user is uncertain about her information need [17]. Furthermore, the information need can shape throughout the search session. For instance, when the user prepares for writing a summary about a particular topic, the search process typically takes several iterations in which the user directs the search by tuning the initial query and the initial search intent, after observing the results. This important type of a search scenario is called *exploratory search* [10, 12, 18].

There is a wide variety of qualitative definitions of exploratory search [19]. Marchionini [12] illustrated exploratory and lookup tasks as an overlapping cloud and suggested that lookup tasks are embedded in exploratory tasks and vice versa. The problem context that motivates the search process is typically characterized in definitions of exploratory search [18]. Imprecise task requirements or open-ended search goals are the two primary attributes often used to define exploratory search with respect to the problem context [11]. The exploratory search process is considered to be cognitively complex with the information seeker being uncertain about the search process [18].

To help the user in exploratory search, her interactions with the system can be employed to infer her search intent. This is challenging for two reasons. First, active interaction is required from the user; however, users are often not willing to invest in actively giving feedback to search systems. Second, even if the interface was appealing enough, the user can only provide a limited amount of feedback, which makes user intent modelling challenging. In this paper we make it easier for the user to give feedback in exploratory search, by allowing feedback on multiple domains, in this case keywords and documents. For this we formulate the user intent modelling task as a new coupled multi-armed bandit problem, instead of using only one multi-armed bandit for one modality as in earlier approaches [14].

The rest of the paper is organized as follows: First we model exploratory search as a learning problem with limited feedback. Next, we propose the *coupled multi-armed bandits* algorithm that employs Thompson sampling with a novel probabilistic user model. We conclude the paper by evaluating the proposed method in a simulation scenario and a user study.

PROPOSED APPROACH

Problem Setting
Let D be a set of documents in a corpus and K be a set of keywords extracted from these documents. For each user, it is assumed that the relevance of each document $d \in D$ is an unknown distribution over [0,1]. This distribution encodes the uncertainty of the user about the relevance of each item, which is a key element of exploratory search. The expected

IUI'16, March 07–10, 2016, Sonoma, CA, USA.
 ISBN 978-1-4503-4137-0/16/03$15.00
DOI: http://dx.doi.org/10.1145/2856767.2856803

relevance of d for the user is denoted by $E_D[d]$. The document $d \in D$ is more relevant than $d' \in D$, if $E_D[d'] < E_D[d]$. Similarly, it is assumed that the relevance of each keyword $k \in K$ is an unknown distribution over [0,1] with its expected relevance denoted by $E_K[k]$. We call this set of distributions the user intent model. The expected relevances of keywords and documents are connected through a model of the data that defines how keywords belong to documents.

In each session, a user with a fixed but unknown intent model arrives. Each session consists of N iterations. In each iteration, the user provides input based on her intent model as relevance values to keywords and documents, and the algorithm provides a set of new documents and keywords. It is assumed that user feedback consists of samples from the relevance distributions.

The retrieval system should look for the most relevant document $d^* = \arg\max_{d \in D} E_D[d]$ and present it to the user. To solve this maximization the system needs to explore the document space to estimate the expected relevance of documents based on user feedback. At the same time it should exploit the estimates to show relevant documents as early as possible. This kind of a black box optimization problem, where the objective function is unknown and expensive to sample, has been studied in multi-armed bandit [6] and Bayesian optimization [4, 16] literature. A natural performance criterion for these problems is regret, which is the loss due to not presenting the most relevant documents to the user. The cumulative regret after receiving feedback for a set of documents n_D is $cum\_regret = |n_D| E_D[d^*] - \sum_{d \in n_D} E_D[d]$. The goal is to minimize the cumulative regret, which is equivalent to maximizing the sum of expected relevance feedback on documents in n_D. However, since the expected relevance of documents is hidden to the system, this measure cannot be calculated in practice.

In practice, an exploratory search system is successful if it presents to the user items, e.g. documents, that the user finds interesting. How we measure this "interest" is through maximizing the average number of clicks (or other types of positive relevance feedback) on documents in N iterations [15]. Since it is reasonable to assume that the user provides positive feedback mostly on relevant documents, this user behavior model also minimizes the cumulative regret defined from the theory point of view.

By modeling the problem as regret minimization, there is little hope to achieve a reasonable result by only considering the limited feedback on documents. In exploratory search, it is more convenient for the user to express the abstract understanding of her needs in terms of higher-level information, such as a set of keywords. In this paper we take advantage of both feedback on documents and on keywords in order to improve the regret of an exploratory search system. We tackle the problem by modeling it as coupled multi-armed bandits where it is possible to provide feedback both to the arms and the features defining the arms. In our exploratory search problem, the arms are documents and the features are keywords defining the documents.

Connecting Documents and Keywords

We assume there exists a document-keyword matrix M defined as

$$M = \begin{bmatrix} P(k_1|d_1) & \cdots & P(k_{|K|}|d_1) \\ \vdots & \ddots & \vdots \\ P(k_1|d_{|D|}) & \cdots & P(k_{|K|}|d_{|D|}) \end{bmatrix}_{|D| \times |K|}$$

where $P(k_i|d_j)$ specifies the likelihood of document d_j generating keyword k_i. This matrix is generated from the data model that expresses how keywords and the documents are related. An example, which we use, is to consider documents as bags of words and to use normalized *tf-idf* representations of documents. We make the simplifying assumption that the connection between expected relevance of a document and keywords is as follows:

$$E_D[d_j] = \sum_{i=1}^{|K|} E_K[k_i] P(k_i|d_j)$$

With the compact notation $E_D[D] = \left[E_D[d_1], \dots, E_D[d_{|D|}]\right]^T$, and analogously for keywords, this becomes

$$E_D[D] = M E_K[K] \quad (1)$$

Note that we made the simplifying assumptions only for the expected relevances; the shapes of the relevance distributions can be different.

Coupled Bayesian Bandits

The user provides relevance feedback to both documents and keywords. The relevance (reward) distributions for document d and keyword k are denoted by $r_d \sim f^D(.|x_d, \theta)$ and $r_k \sim f^K(.|x_k, \theta)$, respectively. The x_d and x_k are the feature (context) vectors associated with d and k. The fixed but unknown parameter θ defines the shared link between these two relevance distributions. After the user has interacted with a set of documents n_D and a set of keywords n_K, we can write the posterior at time $t = |n_D| + |n_K|$ as

$$\pi_t(\theta) \propto \pi_0(\theta) \prod_{d \in n_D} f^D(r_d|x_d, \theta) \prod_{k \in n_K} f^K(r_k|x_k, \theta), \quad (2)$$

where $\pi_0(\theta)$ is the prior distribution for θ. In order to apply Bayesian bandit methods [1, 9] it is only necessary to be able to perform the following two steps: draw a sample from the posterior at time t, and after that score all the documents and keywords by $E_D[r_d|x_d, \theta^p]$ and $E_K[r_k|x_k, \theta^p]$.

These two steps are the minimum requirements for employing the Thompson sampling algorithm for bandits [1, 7, 9]. In Thompson sampling, the exploration and exploitation are controlled indirectly by the uncertainty in the posterior. We can easily draw a sample from the posterior by using any sampling method, after specifying the shared parameter θ that connects the relevance distributions, and the feature vectors x_d and x_k. These parameters are defined by the user model.

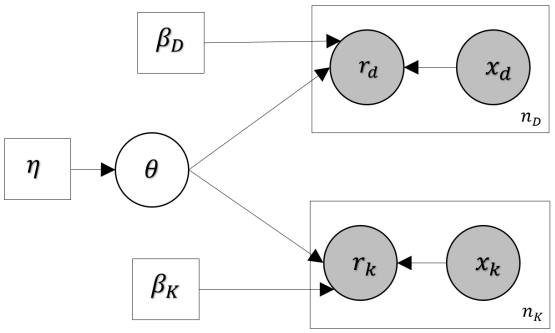

Figure 1. Probabilistic model for user feedback on documents and keywords.

Probabilistic User Model

We propose a simple model for received relevance feedback on documents and keywords. Since the amount of feedback from the user is limited, we need to impose a structure on the expected relevance of items to be able to generalize well. We assume that the expected relevance of keywords is linearly related to their feature vectors by the unknown weight vector θ, i.e. $E_K[K] = M^T\theta$. Based on this linearity assumption and our previous assumption in equation (1), we have $E_D[D] = ME_K[K] = MM^T\theta$. Using this feature transformation, we only need to estimate one set of unknown weights to specify expected relevance of both documents and keywords. Considering Gaussian distributions for relevance, we propose the following model (plate diagram in Figure 1):

$$f^K(r_k|x_k,\theta,\beta_K) = N(r_k; x_k^T\theta, \beta_K^2)$$

$$f^D(r_d|x_d,\theta,\beta_D) = N(r_d; x_d^T\theta, \beta_D^2)$$

$$\pi_0(\theta) = N(\theta; 0, \eta^2\mathrm{I})$$

Here, x_k is the k^{th} column of M and x_d is the d^{th} column of MM^T; they define feature vectors for keyword k and document d, respectively. Since all the distributions are Gaussian, the posterior is also a Gaussian distribution, with:

$$\pi_t(\theta) = N(\theta; \mu_t, \Sigma_t)$$

$$\Sigma_t^{-1} = \beta_D^{-2} X_{n_D}^T X_{n_D} + \beta_K^{-2} X_{n_K}^T X_{n_K} + \eta^{-2}\mathrm{I} \qquad (3)$$

$$\mu_t = \Sigma_t\left(\beta_D^{-2} X_{n_D}^T R_{n_D} + \beta_K^{-2} X_{n_K}^T R_{n_K}\right)$$

where X_{n_D} is a $|n_D| \times D$ design matrix containing feature vectors for the observed documents in the set n_D, and R_{n_D} is a $|n_D| \times 1$ matrix of the corresponding observed relevance values. With the same logic, X_{n_K} is a $|n_K| \times D$ design matrix containing feature vectors for the observed keywords in the set n_K, and R_{n_K} is a $|n_K| \times 1$ matrix of the corresponding observed relevance values. The computational complexity of (3) comes from the covariance matrix inversion.

The coupled multi-armed bandits algorithm employing Thompson sampling for controlling exploration-exploitation tradeoff is as follows:

ALGORITHM 1: Coupled multi-armed bandits

At time step t

1. draw $\theta^p \sim \pi_t(\theta)$ from (3)
2. for document bandit: select $d^+ = \arg\max_{d \in D} x_d^T\theta^p$
3. for keyword bandit: select $k^+ = \arg\max_{k \in K} x_k^T\theta^p$
4. update the posterior (3) based on user feedback and observed feature vectors

SIMULATION

In each exploratory search session, the simulator considers a small set of documents and keywords as targets, and assumes their expected relevance to be 1. The simulated user employs Latent Semantic Indexing (LSI) and cosine similarity measure to calculate similarities to the targets for all documents and keywords. According to these similarities, an expected relevance value in [0,1] is assigned to each document and keyword, defining $E_D[D]$ and $E_K[K]$.

In each iteration, the algorithm presents five documents and five keywords to the user (by repeating steps 2 and 3 of Algorithm 1). We assume that the simulated user selects document d as relevant with probability equal to relevance value $r_d \sim N(E_D[d], \beta_D^2)$, and provides relevance feedback $r_k \sim N(E_K[k], \beta_K^2)$ for keyword k with probability proportional to $|r_k - 0.5|$. The motivation is that users usually give feedback to keywords that are highly relevant or irrelevant. We used a fixed pool of 750 computer science arXiv articles, with only 50 of them having high expected relevance in the user intent model. The model parameters were set to $\beta_D = \beta_k = 0.3$ and $\eta = 0.5$. We compared our method with variants that only consider feedback on keywords or documents to update the posterior (3).

Based on the simulation result in Figure 2, we can conclude that considering feedback on both documents and keywords will significantly increase the number of relevant items that the user can find with the same number of iterations.

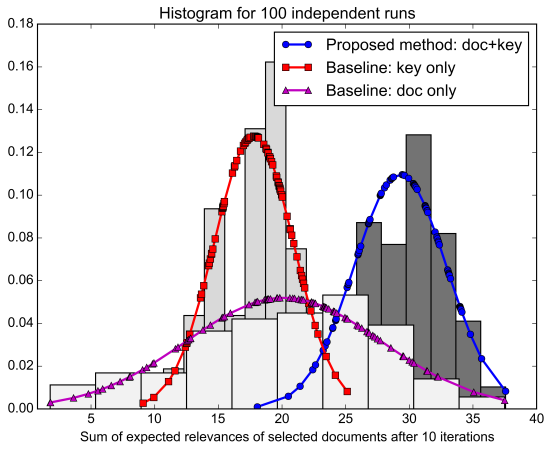

Figure 2. Histogram for 100 independent runs of the methods. The sum of expected relevances of selected documents after 10 iterations is significantly higher when both types of feedback (doc+key) are considered.

USER EXPERIMENTS AND RESULTS

We implemented the proposed method in an existing exploratory search system SciNet [8]. SciNet uses the LinRel algorithm [2] to learn the user intent model based on feedback on keywords visualized on a radar-like interface. In our implementation, we replaced LinRel with Algorithm 1. Furthermore, user interactions with documents, i.e. clicks or bookmarks, are considered as relevance feedback on documents, in addition to feedback on keywords. The computational complexity of both algorithms is the same.

We conducted a preliminary user study on 10 university students and researchers. All the participants were fluent in English and had some background in computer science or a related field. Each participant performed two exploratory tasks in which they had to do a literature survey on a topic and answer three questions in a fixed amount of time. The topics were reinforcement learning and neural networks. The participants reported their knowledge of the topics on 1-5 Likert scale (with 2.2 on average). The user interface and real time performance of the systems were identical, and the participants were naïve about which exploratory search engine was used for each task. In each iteration, five documents were shown to the user and the user could bookmark the relevant ones. All user interactions were logged by the system including the typed query, documents and keywords presented to the user, and documents and keywords that the user interacted with. After each task the participants answered a questionnaire containing SUS [5] and a short version of ResQue [13] using 1-5 Likert scale, where 1 indicates "strongly disagree". A short interview with each participant was conducted after the tasks.

Answers to all the tasks and the bookmarked documents were rated by experts in a double-blind assessment using a scale from 0 (no answer) to 5 (perfect answer). All the shown documents were assessed on a binary scale based on their relevance to the search topic. The inter-rater agreement between the experts showed that the rankings overlapped more than 80%. One of the users was excluded from the analysis because he did not bookmark any article.

Table 1 summarises different performance measures for the proposed algorithm and the baseline system. The percentage of all the shown articles that were labelled as relevant by the experts was calculated as a measure of the quality of the shown information. For the proposed system this value was 5 percent better compared to the baseline. However, the difference was not statistically significant.

Performance measure	Proposed	Baseline
Average SUS score	**75**	67.2
Average ResQue score	**54.7**	52.3
% of relevant documents	**84.6**	79.1
User task performance	3.45	3.45

Table 1. Performance measures for the proposed and baseline systems. The better value on each row is shown in bold.

The proposed algorithm had better SUS and ResQue scores compared to the baseline. However, due to the small sample size these differences were not statistically significant. It should be noted that most of the questions in SUS and ResQue target areas such as user interface, which was the same in both systems. There are two questions in ResQue that measure the *novelty* and *diversity* of search results [13]: *The recommender system helped me discover new items* and *The items recommended to me are diverse*. In both questions the proposed method scored higher (4.1 and 3.7) against the baseline system (3.5 and 3.3).

User task performance was measured by averaging the expert assessments of the answers for the three questions in each task. The users were hard pressed to gather a report in time, which was also evident in their answers. In these reports both systems achieved the same average performance.

In the interviews, 6 out of 9 users reported higher satisfaction with the proposed system (more diversity and better results), one reported that he could not tell the difference, and two said that the baseline was better.

Overall, the preliminary results indicate that the proposed system gave the users a more satisfying image of the topic they were exploring. Furthermore, the usability of the total system has also improved.

DISCUSSION AND CONCLUSION

In this paper we introduced the *coupled multi-armed bandits* algorithm as the exploratory search method that employs the user feedback on both the retrieved items and their features. Our approach is based on two main ideas. First, we model user behavior as a generative probabilistic model. Second, we couple different sources of feedback in a unified model. Our simulation results and preliminary user study indicate that considering these two sources of feedback can improve the performance and quality of the exploratory search.

From the practical point of view, our algorithm provides the opportunity to exploit several types of relevance feedback that are available for documents, in addition to relevance feedback available for individual keywords. For example in [3] it was shown that it is possible to detect implicit relevance feedback from physiological signals such as electrodermal activity and facial electromyography on documents. We believe that this is an important step for the future of exploratory search applications since an increasing amount of research studies the feasibility of performing information retrieval based on novel types of relevance signals both on keywords and on documents.

ACKNOWLEDGMENTS

This work has been partly supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN), Re:Know funded by TEKES, and MindSee (FP7–ICT; Grant Agreement #611570).

REFERENCES

1. Agrawal, S., and Goyal, N. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proc. of ICML,* (2013), 127-135.
2. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, (2003), 397-422.
3. Barral, O., Eugster, M. J., Ruotsalo, T., Spapé, M. M., Kosunen, I., Ravaja, N., ... and Jacucci, G. Exploring Peripheral Physiology as a Predictor of Perceived Relevance in Information Retrieval. In *Proc. of IUI,* ACM (2015), 389-399.
4. Brochu, E., Cora, V. M., and De Freitas, N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.
5. Brooke, J. SUS-A quick and dirty usability scale. *Usability evaluation in industry 189*, (1996), 4-7.
6. Bubeck, S., and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, (2012), 5(1), 1-122.
7. Chapelle, O., and Li, L. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems,* (2011), 2249-2257.
8. Glowacka, D., Ruotsalo, T., Konuyshkova, K., Kaski, S., and Jacucci, G. Directing exploratory search: Reinforcement learning from user interactions with keywords. In *Proc. of IUI,* ACM (2013), 117-128.
9. Hoffman, M. D., Shahriari, B., and de Freitas, N. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Proc. of AISTATS,* (2014), 365-374.
10. Kammerer, Y., Nairn, R., Pirolli, P. and Chi, E.H. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2009), 625-634.
11. Kim, J. Describing and predicting information-seeking behavior on the web. *Journal of the American Society for Information Science and Technology*, (2009), 679–693.
12. Marchionini, G. Exploratory search: from finding to understanding. *Communications of the ACM* (2006), 41–46.
13. Pu, P., Chen, L., and Hu, R. A user-centric evaluation framework for recommender systems. In *Proc. of RecSys,* ACM (2011), 157-164.
14. Ruotsalo, T., Jacucci, G., Myllymäki, P. and Kaski, S. Interactive intent modeling: Information discovery beyond search. *Communications of the ACM*, (2014), 58(1), 86-92.
15. Slivkins, A., Radlinski, F., and Gollapudi, S. Ranked bandits in metric spaces: learning diverse rankings over large document collections. *The Journal of Machine Learning Research*, (2013), 14(1), 399-436.
16. Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. of ICML,* (2010).
17. Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proc. of CHI,* ACM (2004), 415-422.
18. White, R. W., and Roth, R. A. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, (2009), 1-98.
19. Wildemuth, B. M., and Freund, L. Assigning search tasks designed to elicit exploratory search behaviors. In *Proc. of HCIR*. ACM (2012), 1–10

Publication II

Pedram Daee*, Tomi Peltola*, Marta Soare*, and Samuel Kaski. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning*, 106, 9-10, 1599–1620, 2017.

Mach Learn
DOI 10.1007/s10994-017-5651-7

Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction

Pedram Daee[1] · **Tomi Peltola**[1] · **Marta Soare**[1] · **Samuel Kaski**[1]

Received: 5 February 2017 / Accepted: 20 June 2017

Abstract Prediction in a small-sized sample with a large number of covariates, the "small n, large p" problem, is challenging. This setting is encountered in multiple applications, such as in precision medicine, where obtaining additional data can be extremely costly or even impossible, and extensive research effort has recently been dedicated to finding principled solutions for accurate prediction. However, a valuable source of additional information, domain experts, has not yet been efficiently exploited. We formulate knowledge elicitation generally as a probabilistic inference process, where expert knowledge is sequentially queried to improve predictions. In the specific case of sparse linear regression, where we assume the expert has knowledge about the relevance of the covariates, or of values of the regression coefficients, we propose an algorithm and computational approximation for fast and efficient interaction, which sequentially identifies the most informative features on which to query expert knowledge. Evaluations of the proposed method in experiments with simulated and real users show improved prediction accuracy already with a small effort from the expert.

Keywords Bayesian methods · Experimental design · Human-to-machine transfer learning · Interactive machine learning · Statistics in high dimensions

1 Introduction

Datasets with a small number of samples n and a large number of variables p are nowadays common. Statistical learning, for example regression, in these kinds of problems is ill-posed,

Pedram Daee, Tomi Peltola and Marta Soare have contributed equally to this work.

Editors: Kurt Driessens, Dragi Kocev, Marko Robnik-Šikonja, Myra Spiliopoulou.

Samuel Kaski
Samuel.Kaski@aalto.fi

[1] Helsinki Institute for Information Technology HIIT and Department of Computer Science, Aalto University, Konemiehentie 2, Espoo, Finland

Published online: 12 July 2017

and it is known that statistical methods have limits in how low in sample size they can go (Donoho and Tanner 2009). A lot of recent research in statistical methodology has focused on finding different kinds of solutions via well-motivated trade-offs in model flexibility and bias. These include strong assumptions about the model family, such as linearity, low rank, sparsity, meta-analysis and transfer learning from related datasets, efficient collection of new data via active learning, and, less prominently, prior elicitation.

There is, however, a certain disconnect between the development of state-of-the-art statistical methods and their application in challenging data analysis problems. Many applications have significant amounts of previous knowledge to incorporate into the analysis, but this is often unstructured and tacit. Building it into the analysis would require tailoring the model and eliciting the knowledge in a suitable format for the analysis, which would be burdensome for both experts in statistical methods and experts in the problem domain. More commonly, new methods are developed to work well in some broad class of problems and data, and domain experts use default approaches and apply their previous knowledge post-hoc for interpretation and discussion. Even when experts in both fields are directly collaborating, the feedback loop between the method development and application is often slow.

We propose to directly integrate the expert into the modelling loop by formulating knowledge elicitation as a probabilistic inference process. We study a specific case of sparse linear regression with the aim of solving prediction problems where the number of available samples (*training data*) is insufficient for statistically accurate prediction. A core characteristic of the formulation is that it adapts to the feedback obtained from the expert and it sequentially integrates every piece of information before deciding on the next query for the expert. In particular, the predictive regression model and the feedback model are subsumed into a joint probabilistic model, the related uncertainties of which can be sequentially updated, after each expert interaction. The query selection is then naturally formulated as an *experimental design problem*, aiming at maximizing the information gained from the expert in a limited number of queries. This efficiently reduces the burden on the expert, since the most informative queries will be asked first, redundant queries can be avoided via the sequential updating, and the expert's effort is not wasted on aspects of the model, where the training data already provides strong information. Notably, compared to pure prior elicitation, the reduction in the number of interactions makes knowledge elicitation for high-dimensional parameters (such as the regression weights in large p models) practicable. This paper contributes to the interactive machine learning literature, focusing on probabilistic modelling for interactively eliciting and incorporating expert knowledge. The other important aspect, of designing the user interfaces, will be focused on in future work.

1.1 Contributions and outline

After discussing related work (Sect. 2), we rigorously formulate expert knowledge elicitation as a probabilistic inference process (Sect. 3). We study a specific case of sparse linear regression, and in particular, consider cases where the expert has knowledge about the relevance of the covariates or the values of the regression coefficients (Sect. 4). We present an algorithm for efficient interactive sequential knowledge elicitation for high-dimensional models that makes knowledge elicitation in "small n, large p" problems feasible (Sect. 4.3). We describe an efficient computational approach using deterministic posterior approximations allowing real-time interaction for the sparse linear regression case (Sect. 4.4). Simulation studies are presented to demonstrate the performance and to gain insight into the behaviour of the approach (Sect. 5). Finally, we demonstrate that real users are able to improve the predictive performance of sparse linear regression in a proof-of-concept experiment (Sect. 5.4).

2 Related work

The problem we study relates to several topics studied in the literature, either by the method, goal, or by the considered setting. In this section, we highlight the main connections.

2.1 Interactive learning

Interactive machine learning includes a variety of ways to employ user's knowledge, preferences, and human cognition to enhance statistical learning (Ware et al. 2001; Fails and Olsen 2003; Amershi 2012; Robert et al. 2016). These methods have been used successfully in several applications, such as learning user intent (Ruotsalo et al. 2014) and preferential clustering. For instance, the semi-supervised clustering method in Lu and Leen (2007); Balcan and Blum (2008) uses feedback on pairs of items that should or should not be in the same cluster, to learn user preferences. In addition to the differences coming from the learning task, one notable contrast between these works and our method is that their aim is to identify user preferences or opinions, whereas our goal is to use expert knowledge as an additional source of information for an improved prediction model, by integrating it with the knowledge coming from the (small n) data. As a probabilistic approach, our work relates to Cano et al. (2011) and House et al. (2015), where expert feedback is used for improved learning of Bayesian networks and for visual data exploration, respectively. In Sect. 3.3, we show how these works can be seen as instances of the general approach we propose.

2.2 Active learning and experimental design

The method we propose for efficiently using expert feedback is related to active learning techniques [see, for instance, Settles (2010)], where the algorithms actively select the most informative data points to be used in prediction tasks. Our method similarly queries the expert for information with the goal of maximising the information gain from each feedback and thus learning more accurate models with less feedback. The same definition of efficiency with respect to the use of samples also connects our work with experimental design techniques (Kiefer and Wolfowitz 1959; Chaloner and Verdinelli 1995), which considers designing informative experiments for collecting data in settings with limited resources. This has been recently considered in sparse linear settings by Seeger (2008), Hernández-Lobato et al. (2013), Ravi et al. (2016) and is important in many application fields (Busby 2009; Ferreira and Gamerman 2015; Martino et al. 2017). Our task, however, is different as we do not aim at collecting new data samples, but the additional information comes from a different source, the expert, with its respective bias and uncertainty. Indeed, our method will be most useful in cases where obtaining additional input samples would be too expensive. Active learning has also been used to query feature labels rather than new data points in natural language processing applications (Druck et al. 2009; Settles 2011; Raghavan et al. 2006). The difference between these works and our paper comes from the task (they consider classification rather than prediction), the model assumptions (they did not consider sparse models which are suitable for "small n, large p" settings), and the feedback type.

2.3 Prior elicitation

Many works have studied approaches for efficient elicitation of expert knowledge. Typically, the goal of prior elicitation techniques (O'Hagan et al. 2006) is to use expert knowledge to construct a prior distribution for Bayesian data analysis and restrict the range of parameters

to be later used in learning models. In particular, Garthwaite and Dickey (1988), Kadane et al. (1980) study methods of quantifying subjective opinion about the coefficients of linear regression models through the assessment of credible intervals. These elicitation methods were shown to obtain prior distributions that represent well the expert's opinion. Similar elicitation methods have been employed in a wide range of application settings, in which expert knowledge elicitation techniques have been studied, for instance *preference model elicitation* (Azari Soufiani et al. 2013), or *software development processes* (Hickey and Davis 2003). Our approach goes beyond pure prior elicitation as the training data is used to facilitate efficient user interaction. The concurrent works by Micallef et al. (2017), Afrabandpey et al. (2016), and Soare et al. (2016) also use elicited expert knowledge for improving prediction. Micallef et al. (2017) use a separate multi-armed bandit model to facilitate the elicitation and directly modify the regression model priors, Soare et al. (2016) consider predictions for a target patient in a simulated user setting, while Afrabandpey et al. (2016) consider pairwise similarity feedback on features and uses those to create a better covariance matrix for the prior of coefficients of ridge regression. Contrary to our work, these works do not formulate an encompassing probabilistic model for the expert knowledge and prediction, a crucial feature of our approach. Moreover, they do not consider a sparse regression model and the type of feedback is different.

3 Knowledge elicitation as interactive probabilistic modelling

In the following, we formulate expert knowledge elicitation as a probabilistic inference process.

3.1 Key components

Let y and x denote the outputs (target variables) and inputs (covariates), and θ and ϕ_y the model parameters. Let f encode input (*feedback*) from the user, presumably a domain expert, and ϕ_f be model parameters related to the user input. We identify the following key components:

1. An observation model $p(y \mid x, \theta, \phi_y)$ for y.
2. A feedback model $p(f \mid \theta, \phi_f)$ for the expert's knowledge.
3. A prior model $p(\theta, \phi_y, \phi_f)$ completing the hierarchical model description.
4. A query algorithm and user interface that facilitate gathering f iteratively from the expert.
5. Update process of the model after user interaction.

The observation model can be any appropriate probability model. It is assumed that there is some parameter θ, possibly high-dimensional, that the expert has knowledge about. The expert's knowledge is encoded as (possibly partial) feedback f that is transformed into information about θ via the feedback model. Of course, there could be a more complex hierarchy tying the observation and feedback models, and the feedback model can also be used to model more user-centric issues, such as the quality of or uncertainty in the knowledge or user's interests.

The feedback model, together with a query algorithm and a user interface, is used to facilitate an efficient interaction with the expert. The term "query algorithm" is used here in a broad sense to describe any mechanism that is used to intelligently guide the user's focus in providing feedback to the system. This enables considering a high-dimensional f without overwhelming the expert as the most useful feedbacks can be queried first. Crucially,

this enables going beyond pure prior elicitation as the observed data can be used to inform the queries via the dependence of the feedback and observation models. For example, the queries can be formed as solutions to decision or experimental design tasks that maximize the expected information gain from the interaction.

Finally, as the expert feedback is modelled as additional data, Bayes theorem can be used to sequentially update the model during the interaction. For real-time interaction, this may present a challenge as computation in probabilistic models can be demanding. It is known that slow computation can impair effective interaction (Fails and Olsen 2003) and, thus, efficient computational approaches are important.

3.2 Overall interaction scheme

Figure 1 depicts the information flow. First, the posterior distribution $p(\theta|\mathcal{D})$ given the observations $\mathcal{D} = \{(y_i, x_i) : i = 1, \ldots, n\}$ is computed. Then, the expert is queried iteratively for feedback via the user interface and the query algorithm. The feedback is used to sequentially update the posterior distribution. The query algorithm has access to the latest beliefs about the model parameters and the predicted user behaviour, that is, the posterior predictive distribution of f, $p(f_{t+1} \mid \mathcal{D}, f_1, \ldots, f_t)$, where the f_j are possibly partial observations of f. Based on this information, the query algorithm chooses the most informative queries, or more generally interactions in the user interface.

3.3 Examples

The goal in this paper is to use the interaction scheme (Fig. 1) to help solve prediction problems in the "small n, large p" setting. The approach as described above is, however, more general and applicable to other problems as well. We briefly describe two earlier works that can be seen as instances of it.

Cano et al. (2011) present a method for integrating expert knowledge into learning of Bayesian networks. The observation model is a multinomial Bayesian network with Dirichlet priors. The expert provides answers to queries about the presence or absence of edges in the network and the feedback model assumes the answers to be correct with some probability. Which edge to query about next is selected by maximising the information gain with regard to the inclusion probability of the edges. Monte Carlo algorithms are used for the computation.

House et al. (2015) present a framework for interactive visual data exploration. They describe two observation models, principal component analysis and multidimensional scal-

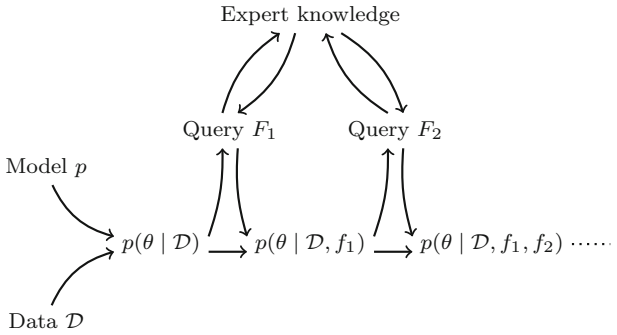

Fig. 1 Information flow. The parameters ϕ_y and ϕ_f are omitted from the posterior distributions for brevity

ing, that are used for dimensionality reduction to visualise the observations in a two dimensional plot. They do not have a query algorithm, but their user interface allows moving points in a low-dimensional plot closer or further apart. A feedback model then transforms the feedback into appropriate changes in the parameters shared with the observation model to allow exploration of different aspects of the data. Their model affords closed form updates.

4 Feedback models and query algorithm for sparse linear regression

We next introduce the knowledge elicitation approach for sparse linear regression.

4.1 Sparse regression model

Let $\boldsymbol{y} \in \mathbb{R}^n$ be the observed output values and $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ the matrix of covariate values. We assume a Gaussian observation model for the regression[1]:

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I}),$$

where $\boldsymbol{w} \in \mathbb{R}^m$ are the regression coefficients and σ^2 is the residual variance. We assume a gamma prior on the inverse of σ^2 (that is, residual precision; or equivalently, inverse-gamma prior on the variance):

$$\sigma^{-2} \sim \mathrm{Gamma}(\alpha_\sigma, \beta_\sigma).$$

Other tasks, such as classification, could be accommodated by changing the assumption about the observation model to an appropriate generalized linear model.

A sparsity-inducing spike-and-slab prior (George and McCulloch 1993) is put on the regression coefficients $\boldsymbol{w}$:

$$\begin{aligned} w_j &\sim \gamma_j \,\mathrm{N}(0, \psi^2) + (1 - \gamma_j)\delta_0, \quad && j = 1, \ldots, m, \\ \gamma_j &\sim \mathrm{Bernoulli}(\rho), && j = 1, \ldots, m, \end{aligned}$$

where the γ_j are latent binary variables indicating inclusion or exclusion of the covariates in the regression. For covariates *included* in the model, $\gamma_j = 1$ and w_j is drawn from a zero-mean Gaussian distribution with variance ψ^2. For covariates *excluded* from the model $\gamma_j = 0$ and $w_j = 0$ via the Dirac delta point mass at zero, δ_0. We will also refer to covariates included in the model as *relevant* for the regression and covariates excluded as *not-relevant*. The prior inclusion probability of the covariates ρ controls the expected number of covariates included (i.e., the sparsity of model). The α_σ, β_σ, ψ^2, and ρ are fixed hyperparameters.

After observing a *training dataset* $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, the posterior distribution of the regression model is computed using the Bayes theorem[2] as

$$p(\boldsymbol{w}, \boldsymbol{\gamma}, \sigma^2 \mid \mathcal{D}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\sigma^2) p(\boldsymbol{w} \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma})}{p(\boldsymbol{y} \mid \boldsymbol{X})}.$$

The predictive distribution for a new data point $\tilde{\boldsymbol{x}}$ is

$$p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathcal{D}) = \int p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}, \sigma^2 \mid \mathcal{D}) d(\boldsymbol{w}, \sigma^2).$$

[1] The parametrizations of the distributions follow Gelman et al. (2014, Appendix A).

[2] We use the generic $p(\cdot)$ notation, where it is understood that the parameters identify the particular distribution. See, e.g., Gelman et al. (2014, p. 6).

4.2 Feedback models

Feedback models are used to incorporate the expert knowledge into the regression model. They extend the regression model such that both parts are subsumed into a single probabilistic model, where information flows naturally between the parts (following the Bayesian modelling paradigm).

The feedback model is naturally dependent on the available type of expert knowledge in the targeted application. For our formulation we consider two simple and natural feedback models encoding knowledge about the individual regression coefficients:

- Expert has knowledge about the value of the coefficient ($f_{w,j} \in \mathbb{R}$):

$$f_{w,j} \sim \mathrm{N}(w_j, \omega^2). \tag{1}$$

  We do not assume that the expert can give exact values for the coefficients, but that there is some uncertainty in the expert's estimates, the amount of which is controlled by the variance ω^2. The smaller the ω^2, the more accurate the knowledge is assumed a priori, and the stronger the change in the model in response to the feedback.
- Expert has knowledge about the relevance of coefficient ($f_{\gamma,j} \in \{0, 1\}$ for *not-relevant*, *relevant*):

$$f_{\gamma,j} \sim \gamma_j \,\mathrm{Bernoulli}(\pi) + (1 - \gamma_j)\,\mathrm{Bernoulli}(1 - \pi). \tag{2}$$

  Here, π models uncertainty of the knowledge (akin to ω^2 above). *A priori*, the expert feedback is expected to be 1 (*relevant*) with probability π if $\gamma_j = 1$ (the covariate is *included* in the regression). In other words, π can be thought of as the probability of the expert being correct in his or her feedback relative to the state of the covariate inclusion γ_j.

The posterior distribution after getting a set $\mathcal{F} = (\boldsymbol{f}_{\boldsymbol{w}}, \boldsymbol{f}_{\gamma})$ of expert feedback, where, with some abuse of notation, $\boldsymbol{f}_{\boldsymbol{w}}$ and $\boldsymbol{f}_{\gamma}$ collect the given feedback (not assumed to be necessarily available for all covariates) is

$$p(\boldsymbol{w}, \boldsymbol{\gamma}, \sigma^2 \mid \mathcal{D}, \mathcal{F}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2)p(\sigma^2)p(\boldsymbol{w} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\boldsymbol{f}_{\gamma} \mid \boldsymbol{\gamma})p(\boldsymbol{f}_{\boldsymbol{w}} \mid \boldsymbol{w})}{p(\boldsymbol{y}, \boldsymbol{f}_{\gamma}, \boldsymbol{f}_{\boldsymbol{w}} \mid \boldsymbol{X})}.$$

The predictive distribution follows as before but with the new posterior distribution. The posterior distribution can be updated sequentially when more feedback is collected, as explained in Sect. 3. In the following, we will assume only one type of feedback at a time in the modelling, but this could be extended to multiple simultaneous types.

To relate the work to prior elicitation, we can think of first updating the posterior distribution with only the expert feedback $\mathcal{F}$, and then, using the posterior as the prior distribution for updating the model with the training set observations $\mathcal{D}$. However, incorporating the expert knowledge through the feedback models (instead of directly as priors) is crucial for the sequential knowledge elicitation, as it allows computing the predictive distributions for the feedback. Moreover, the sequential elicitation exploits the information in the training data to facilitate an efficient elicitation process for high-dimensional parameters.

4.3 Query algorithm

Our aim is to improve prediction. Thus, the user interaction should focus on aspects of the model (here, predictive covariates or features; we use the terms interchangeably) that would be most beneficial towards this goal. We use the query algorithm to rank the features for

choosing which one to ask feedback about next. The ranking is formulated as a Bayesian experimental design task (Chaloner and Verdinelli 1995) given all information collected thus far.

The utility function used for scoring the alternative queries can be tailored according to the application. In this paper we use information gain in the prediction, defined as the Kullback–Leibler divergence (KL) between the current posterior predictive distribution $p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathcal{D}, \mathcal{F})$ and the posterior predictive distribution with the new feedback f_j, $p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathcal{D}, \mathcal{F}, f_j)$. The bigger the information gain, the bigger impact the new feedback has on the predictive distribution. More specifically, the feature j^* that maximizes the expected information gain is chosen next:

$$j^* = \arg\max_{j \notin \mathcal{F}} \mathbb{E}_{p(\tilde{f}_j \mid \mathcal{D}, \mathcal{F})} \left[\sum_i \mathrm{KL}[p(\tilde{y} \mid \boldsymbol{x}_i, \mathcal{D}, \mathcal{F}, \tilde{f}_j) \parallel p(\tilde{y} \mid \boldsymbol{x}_i, \mathcal{D}, \mathcal{F})] \right],$$

where j indexes the features, $\mathcal{F}$ is the set of feedbacks that have already been given, and the summation over i goes over the training dataset. Since the feedback itself will only be observed after querying the expert, we take the expectation over the posterior predictive distribution of the feedback $p(\tilde{f}_j \mid \mathcal{D}, \mathcal{F})$. More details about the Bayesian experimental design are provided in Appendix B.

As a side note, if the predictive distribution of y was Gaussian, the problem would be simple. The expected information gain would be independent of y and the actual values of the feedbacks (when feedback is on values of the regression coefficients), and would only depend on the x and on which features the feedback was given (Seeger 2008). The sparsity-promoting prior, however, makes the problem non-trivial.

4.4 Computation

The model does not have a closed-form posterior distribution, predictive distribution, or solution to the information gain maximization problem. To achieve fast computation, we use deterministic posterior approximations. Expectation propagation (Minka 2001) is used to approximate the spike-and-slab prior (Hernández-Lobato et al. 2015) and the feedback models, and variational Bayes (e.g., Bishop 2006, Chapter 10) is used to approximate the residual variance σ^2. The form of the posterior approximation for the regression coefficients $\boldsymbol{w}$ is Gaussian. The posterior predictive distribution for y is also approximated as Gaussian. Details are provided in Appendix A.

Expectation propagation has been found to provide good estimates of uncertainty, which is important in experimental design (Seeger 2008; Hernández-Lobato et al. 2013; Hernández-Lobato et al. 2015). In evaluating the expected information gain for a large number of candidate features, running the approximation iterations to full convergence for each is too slow, however. We follow the approach of Seeger (2008), Hernández-Lobato et al. (2013) in computing only a single iteration of updates on the essential parameters for each candidate. We show in the results that this already provides a good performance for the query algorithm in comparison to random queries. Details on the computations are provided in Appendix B.

Markov chain Monte Carlo (MCMC) methods could alternatively be used for computation, but sampling efficiently over the binary space of size 2^m for $\boldsymbol{\gamma}$ can be difficult (Peltola et al. 2012; Schäfer and Chopin 2013) and naive approaches would be slow. Sequential Monte Carlo (SMC) algorithms (Del Moral et al. 2006; Schäfer and Chopin 2013) are designed to move between distributions that change in steps and could provide a feasible alternative to deterministic approximations here. However, designing efficient SMC algorithms for the

spike and slab model is not trivial (see Schäfer and Chopin (2013) for an approach). We have not evaluated using MCMC computation in this work.

5 Experiments

The performance of the proposed method is evaluated in several "small n, large p" regression problems on both simulated and real data.[3] A proof-of-concept user study is presented to demonstrate the feasibility of the method with real users. We compare our sequential design algorithm (Sect. 4.3) to two baselines in the experiments:

- Random feature suggestion,
- The non-sequential version of our algorithm, which chooses the sequence of features to be queried before observing any expert feedback.

Additionally, to provide a yardstick, we plot the results of an "oracle," which knows the relevant features beforehand, and which is obviously not available in practice:

- Query first on the relevant features, and then choose at random from the features not already selected.

We compare the performance of these strategies with synthetic and real data, and with both simulated and real users providing feedback.

5.1 Synthetic data

We use synthetic data to study the behaviour of the approach in a wide range of controlled settings.

5.1.1 Setting

The covariates of n training data points are generated from $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. Out of the m regression coefficients $w_1, \ldots, w_m \in \mathbb{R}$, m^* are generated from $w_j \sim \mathrm{N}(0, \psi^2)$ and the rest are set to zero. The observed output values are generated from $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2\mathbf{I})$. We consider cases where the user has knowledge about the value of the coefficients (Eq. 1 with noise value $\omega = 0.1$) and where the user has knowledge about whether features are relevant or not (Eq. 2 with $\gamma_j = 1$ if w_j is non-zero, and $\gamma_j = 0$ otherwise, and $\pi = 0.95$). For a generated set of training data, all algorithms query feedback about one feature at a time. Mean squared error (MSE) is used as the performance measure to compare the query algorithms. For the simulated data setting, we use the known data-generating values for the fixed hyperparameters, namely: $\psi^2 = 1,\ \rho = m^*/m,$ and $\sigma^2 = 1$ (instead of the inverse-gamma prior of σ^2 used in the rest of the experiments).

5.1.2 Results

Sequential experimental design requires only few feedbacks to improve the MSE In Fig. 2, we consider a "small n, large p" scenario, with $n = 10, m^* = 10$ and with increasing dimensionality (hence also increasing sparsity) from $m = 12, \ldots, 200$. The heatmaps show the average MSE values over 100 runs (repetitions of the data generation) for both feedback

[3] All codes and data are available in https://github.com/HIIT/knowledge-elicitation-for-linear-regression.

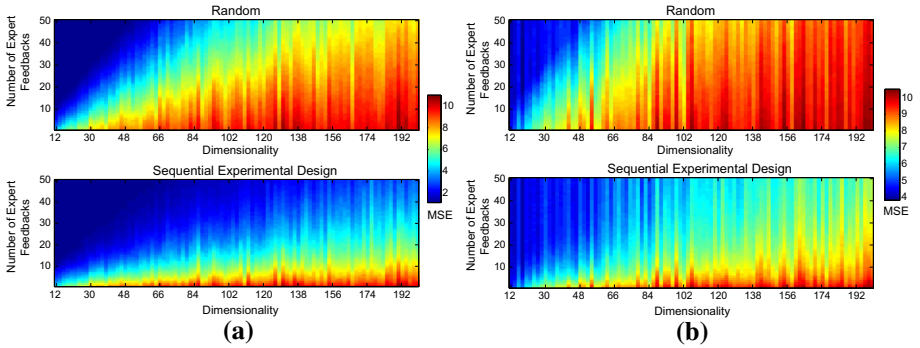

Fig. 2 Mean squared errors in simulated settings with increasing dimensionality. The number of relevant coefficients $m^* = 10$ and the number of training data points $n = 10$. The MSE values are averages over 100 independent runs. **a** Feedback on coefficient values, **b** feedback on coefficient relevances

models, as obtained by our sequential experimental design algorithm and by a strategy that randomly selects the sequence of features on which to ask for expert feedback. The result shows that our method achieves a faster improvement in the prediction, starting from the very first expert feedbacks, for both feedback types, and at all the dimensionalities. Notably, in the case of the random strategy, the performance decreases rapidly with the growing dimensionality (even with 50 feedbacks, in the setting with 200 dimensions, the prediction error for random strategy stays high), while the expert feedback via the sequential experimental design is informative enough to provide good predictions even in large dimensionalities. Comparing the two types of feedback, the feedback on the coefficient values gives better performance for both strategies.

Sequential experimental design requires only few training data points to identify informative queries Figure 3 shows heatmaps for the same setting but with a fixed dimension $m = 100$ and increasing numbers of training data points $n = 5, \ldots, 50$. For very small sample sizes ($n < 10$), a difference between the performance of the two methods starts being visible after 20-30 feedbacks. For larger training samples sizes ($10 < n < 30$), the MSE reduction in our method is more visible from the first feedbacks, while for $n > 30$, both strategies have much smaller MSE after receiving the first feedbacks. Thus, the experiments

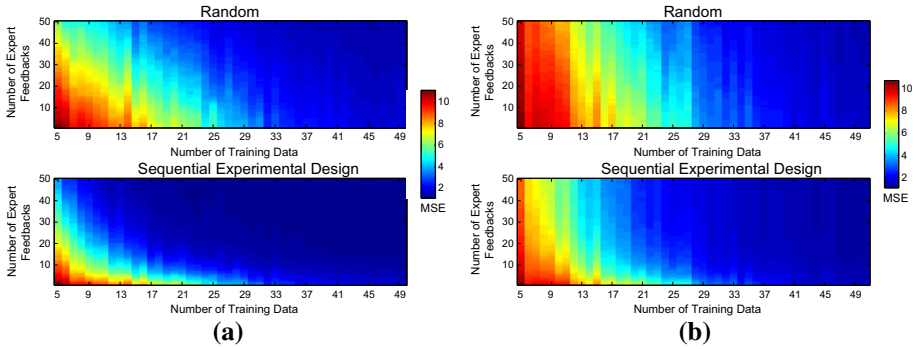

Fig. 3 Mean squared errors as a function of the number of training data points (*horizontal axis*) and number of expert feedbacks (*vertical axis*). The number of relevant coefficients $m^* = 10$ and the number of dimensions $m = 100$. The MSE values are averages over 100 independent runs. **a** Feedback on coefficient values, **b** feedback on coefficient relevances

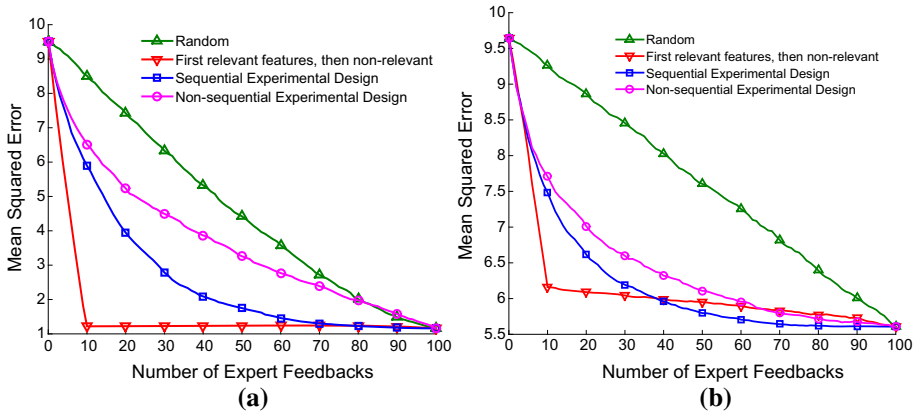

Fig. 4 MSE for all query algorithms, with simulated data, for feedback on coefficient values and relevance. Note that the *red* strategy is not available in practice. **a** Feedback on coefficient values, **b** feedback on coefficient relevances (Color figure online)

show that for all values of n there is an improvement in the prediction error, from the very first expert feedback. An important observation is that the largest improvements come when the training data does not alone provide enough information.

Expert feedback affects the next query We now study the difference between our method and its non-sequential version for the two feedback models. The non-sequential version chooses the sequence of features to be queried before observing any expert feedback. In Fig. 4, we consider a "small n, large p" scenario, with $n = 10$, $m = 100$, $m^* = 10$, and we report the average MSE value over 500 runs. The figure shows that with both experimental design methods the prediction loss decreases rapidly already in the first iterations. In other words, both methods manage to rapidly identify the most informative coefficients and ask about them. This is more evident in the feedback model about coefficient relevance (Fig. 4b). It is also clearly visible that the sequential version is able to reduce the prediction error faster. Also, as expected, the difference between the sequential and non-sequential experimental designs is larger in the case of the stronger feedback model on coefficient values (Fig. 4a).

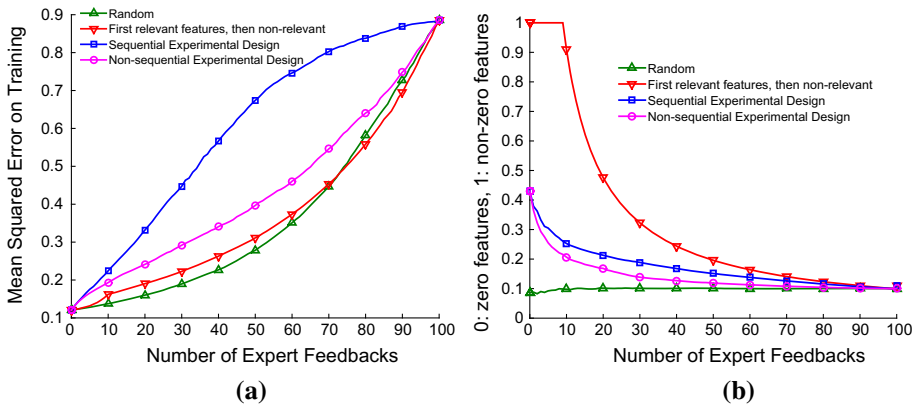

Fig. 5 MSE on the training data and average suggestion behaviour for all query algorithms, with simulated data, for the case where feedback is on coefficient values. **a** MSE on training data, **b** average query behaviour

Expert feedback improves the generalization performance We can get some insight into the behaviour of the approach by comparing the training and test set errors shown in Figs. 4a and 5a for the simulated data scenario described in the previous section with feedback on the coefficient values. The training set error begins to increase as a function of the number of expert feedbacks. This happens because the model without any feedbacks has exhausted the information in the training data (to the extent allowed by the regularizing priors) and fits the training data well. The expert feedback, however, moves the model away from the training data optimum and towards better generalization performance. Indeed, the MSE curves for the training and test errors converge close to each other as the number of feedbacks increases. Moreover, Fig. 5b shows the average (over the replicate runs) query behaviour of each method, indicating whether the methods have queried non-zero features (value 1 in the vertical axis) or zero features (value 0 in the vertical axis). A comparison of Figs. 4a and 5b shows that the convergence is faster for the query algorithms that start by suggesting the non-zero features, implying that these features are more informative.

5.2 Real data: review rating prediction

We test the proposed method in the task of predicting review ratings from textual reviews in subsets of Amazon and Yelp datasets. Each review is one data point, and each distinct word is a feature with the corresponding covariate value given by the number of appearances of the word in the review. Sparse linear regression models have been shown suitable for this task in previous studies, for instance, in Hernández-Lobato et al. (2015).

Amazon data The Amazon data is a subset of the sentiment dataset of Blitzer et al. (2007). This dataset[4] contains textual reviews and their corresponding 1-5 star ratings for Amazon products. Here, we only consider the reviews for products in the *kitchen appliances* category, which amounts to 5149 reviews. The preprocessing of the data follows the method described by Hernández-Lobato et al. (2015), where this dataset was used for testing the performance of a sparse linear regression model. Each review is represented as a vector of features, where the features correspond to unigrams and bigrams, as given by the data provided by Blitzer et al. (2007). For each distinct feature and for each review, we created a matrix of occurrences and only kept for our analysis the features that appeared in at least 100 reviews, that is, 824 features.

Yelp data The second dataset we use is a subset of the YELP (academic) dataset.[5] The dataset contains 2.7 million restaurant reviews with ratings ranging from 1 to 5 stars (rounded to half-stars). Here, we consider the 4086 reviews from the year 2004. Similarly to the preprocessing done for Amazon data, each review is represented as a vector of features (distinct words). After removing non-alphanumeric characters from the words and removing words that appear less than 100 times, we have 465 words for our analysis.

5.3 Simulated expert feedback

For all experiments on Amazon and Yelp datasets, we proceeded as follows: First, each dataset was partitioned in three parts: (1) a training set of 100 randomly selected reviews, (2) a test set of 1000 randomly selected reviews, and (3) the rest as a "user-data set" for

[4] https://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

[5] https://www.yelp.com/dataset_challenge.

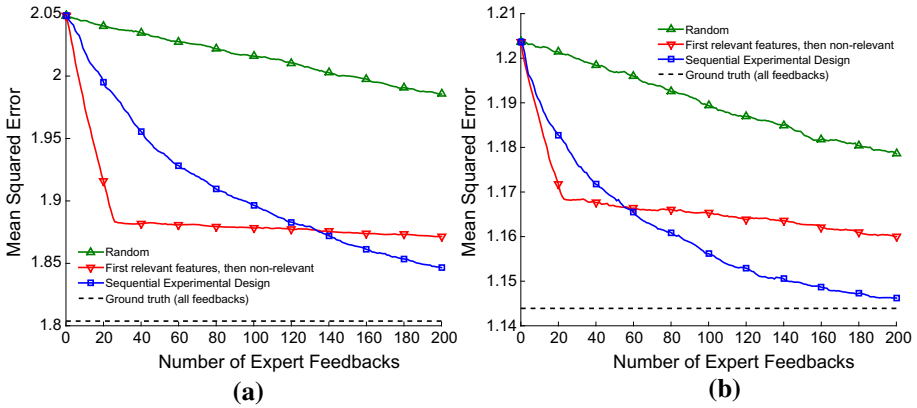

Fig. 6 Mean squared errors when user feedback is on relevance of features for Amazon and Yelp data. The MSE values are averages over 100 independent runs. Note that the *red* strategy is not available in practice. **a** Amazon data, **b** Yelp data (Color figure online)

constructing simulated expert knowledge. The data were normalised to have zero mean and unit standard deviation on the training and user-data sets. The simulated expert feedback was generated based on the posterior inclusion probabilities $\mathbf{E}[\gamma]$ in a spike-and-slab model trained on the user-data partition. We only considered the more realistic case where the expert can give feedback about the relevance of the words. For a word j selected by the algorithm, the expert gives feedback that the word is *relevant* if $\mathbf{E}[\gamma_j] > \pi$, *not-relevant* if $\mathbf{E}[\gamma_j] < 1 - \pi$, and *uncertain* otherwise. The intuition is that if the user-data indicate that a feature is zero/non-zero with high probability, then the simulated expert would select that feature as *not-relevant*/*relevant*. However, for *uncertain* words, the feedback iteration passes without receiving any feedback. For both datasets, the model hyperparameters were set as $\alpha_\sigma = 1$, $\beta_\sigma = 1$, the prediction parameters were tuned based on cross-validation before observing any feedback to $\psi^2 = 0.01$, and $\rho = 0.3$, and the probability that the simulated expert feedback is correct was set to $\pi = 0.9$. All algorithms query feedback about one feature at a time and MSE is used as the performance measure. The ground truth line represents the MSE after receiving expert feedback for all words in each dataset.

5.3.1 MSE improvement with feedback on feature relevances

A first observation of Fig. 6 is that the use of additional knowledge coming from the simulated expert reduces the prediction errors, for all algorithms and on both datasets. The second observation is that the reduction in the prediction error differs significantly depending on whether the methods manage to query feedback on the most informative features first. Indeed, the goal is to make the elicitation as little burdensome as possible for the experts. To reach the goal, a strategy needs to rapidly extract a maximal amount of information from the expert, which here amounts to the careful selection of the features on which to query feedback. As expected, the random query selection strategy has a constant and slow improvement rate, as the number of feedbacks grows, leaving a big gap from the ground-truth performance in both datasets, even after 200 expert feedbacks. If an "oracle" was available to tell which features are relevant, the (unrealistic) strategy to first ask about relevant features would produce a

Table 1 Number of samples/feedbacks needed to reach a particular MSE level in Yelp dataset

MSE	More samples		More feedback	
	Random	Active (Seeger 2008)	Random	SeqExpDes
1.20	21	3	30	3
1.19	55	6	96	11
1.18	94	12	185	25
1.17	146	22	266	46
1.16	241	44	324	85

The values are averages over 100 independent runs

steep increase in performance for the first iterations (26 words for Amazon and 23 for Yelp are marked as relevant, as computed from the full dataset); then it would continue with a very slow improvement rate coming from asking not-relevant words. Our method manages to identify the informative features rapidly and thus has a higher improvement compared to random from the first expert feedbacks. In the case of Yelp data, our strategy manages to be very close to the "oracle" in the initial feedbacks and then converges very close to the ground truth after 200 interactions. Furthermore, there is a significant gap compared to the random strategy for all amounts of feedbacks. In the more difficult (in terms of rating prediction error and size of dimensions) Amazon dataset, the gap to the random strategy is clear but our strategy exceeds the information gain obtained in the 26 non-zero features only after 140 feedbacks.

5.3.2 Expert knowledge elicitation versus collecting more samples

We next contrast the improvements in the predictions brought by eliciting the expert feedback to improvements gained if additional samples could be measured. In this experiment, we do have additional samples, and for choosing them we use two alternative strategies: randomly selecting a sequence of reviews to be included in the training set, and an active learning strategy, which selects samples based on maximizing expected information gain [an adaptation of the method by Seeger (2008)].

Tables 1 and 2 show how many *feedbacks* (for the knowledge elicitation strategies in the last two columns: random and our method; see Sect. 4.3) and respectively how many *additional samples* (that is, additional reviews to be included in the train set) are needed to reach *set levels of MSE*, noting that all strategies have the same "small n, large p" regression setting as a starting point, with $n = 100$.

The number of expert feedbacks required for a given performance level is of the same order of magnitude as the number of additional data needed (Table 1). This slightly surprising finding is even more remarkable since the expert feedback is of a weak type (feedback on the relevance of features). For instance, in Yelp dataset (Table 1) the same level of MSE = 1.18 is obtained either by asking an expert about the relevance of 25 features and by actively selecting 12 extra samples. When the active selection is not possible, we can see that the same information gain requires 94 additional randomly selected samples. Naturally, the results obtained are specific for this Yelp data and for the feedback model we assume. Nevertheless, the comparison shows the potential of expert knowledge elicitation in prediction for settings where getting additional samples is impossible or very expensive. The same observations and intuitions hold for the Amazon data (see Table 2).

Table 2 Number of samples/feedbacks needed to reach a particular MSE level in Amazon dataset

MSE	More samples		More feedback	
	Random	Active (Seeger 2008)	Random	SeqExpDes
2.025	8	4	73	9
2	15	7	152	19
1.975	29	12	>200	31
1.95	44	43	>200	43
1.925	59	71	>200	64
1.9	98	92	>200	95
1.875	>200	144	>200	136

The results are averages over 100 independent runs

5.4 User study

The goal of the user study is to investigate the prediction improvement and convergence speed of the proposed sequential method based on human feedback. Our focus is on testing the accuracy of feedback from real users on the easily interpretable Amazon data rather than on details of the user interface. We asked ten university students and researchers to go through all the 824 words and give us feedback in the form of *not-relevant*, *relevant*, or *uncertain*. This allowed for a fast collection of feedbacks and we could use the pre-given feedback to test the effectiveness of several query algorithms. We assumed that the algorithms had access to 100 training data points and at each iteration they could query the pre-given feedback of the participant about one word. The whole process was repeated for 40 runs, where training data were randomly selected. The hyperparameters of the model were set to the same values as in the simulated data study with the only difference that the strength of user knowledge was lowered to $\pi = 0.7$.

Figure 7 shows the average MSE improvements for each of the ten participants, when using our proposed method and the random query order. From the very first feedbacks, the sequential experimental design approach performs better for all users and captures the expert knowledge more efficiently. The random strategy exhibits a constant rate of performance improvement with increasing number of feedbacks, while the sequential experimental design shows faster improvement rate in the beginning of the interaction. To further quantify the statistical evidence for the difference, we computed the paired-sample t tests between the random suggestion and the proposed method at each iteration (green and blue curves in Fig. 7). Already after the first feedback, the difference between the methods is significant at the Bonferroni corrected level $\alpha = 0.05/200$.

We complement the analysis of the results of the user study with two illustrations. First, to compare the convergence speed of different methods, we normalised the MSE improvements at each iteration by the amount of total improvement obtained by each of the users, when considering all their individual feedback. Figure 8a depicts the convergence speed of methods based on this measure. As can be seen from the figure, for all participants, the proposed method was able to capture most of the participants's knowledge with small budget of feedback queries (stabilizing at around 200 out of the total 824 features in the considered subset of Amazon data). Then, in Fig. 8b, we show the average suggestion behaviour of the methods. One can notice that our algorithm started by favoring queries about relevant words and after exhausting them, the suggestion behaviour moved to querying not-relevant words. The

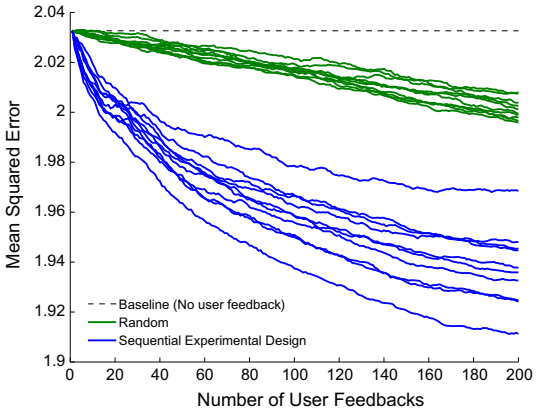

Fig. 7 Mean squared errors for ten participants (average values over 40 independent runs)

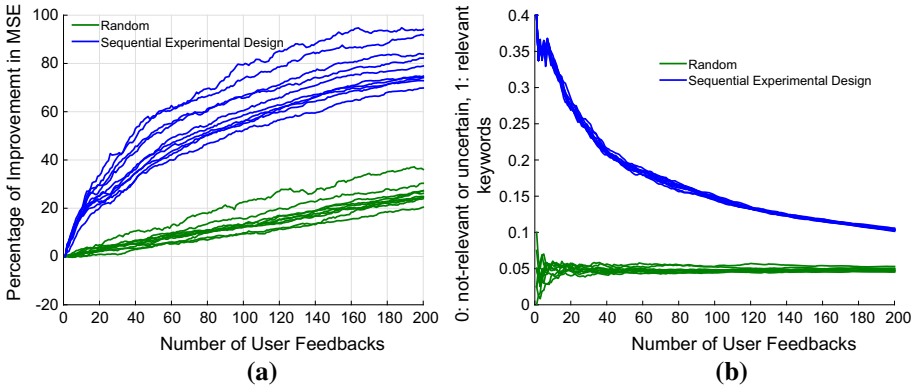

Fig. 8 **a** Convergence speed of different methods to reach the performance achieved by considering all the individual user feedbacks. **b** Average suggestion behaviour of the methods.

relevant words were identified by considering all the data in Amazon dataset and training an spike and slab model and then choosing words with $\mathbf{E}[\gamma_j] > 0.7$ (words with high posterior inclusion probability). Based on this threshold, 39 out of the total of 824 words were considered as relevant.

6 Conclusion and future work

We introduced an interactive knowledge elicitation approach for high-dimensional sparse linear regression. The results for "small n, large p" problems in simulated and real data with simulated and real users, and with expert knowledge on the regression weight values and on the relevance of features, showed improved prediction accuracy already with a small number of user interactions. The knowledge elicitation problem was formulated as a probabilistic inference process that sequentially acquires and integrates expert knowledge with the training data. Compared to pure prior elicitation, the approach can facilitate richer interaction and be used in knowledge elicitation for high-dimensional parameters without overwhelming the expert.

As a by-product of our study, we noticed that even for the rather weak feedback on the relevance of features, the number of expert feedbacks and the number of randomly acquired additional data samples needed to reach a certain level of MSE reduction were of the same order. Although this observation was obtained on a noisy dataset and for a simplifying user interaction setting, the fact that the considered feedback type was rather weak highlights that elicitation from experts is promising.

The presented knowledge elicitation method is widely applicable also beyond the specific assumptions made in this paper. Since all assumptions have been explicated as a probabilistic model, the assumptions can rigorously tailored to match specifics of other data, feedback models, and knowledge elicitation setups, and hence the approach can be applied more generally. The presented results show that it is possible to improve predictions even with preliminary types of feedback, and can be seen as a proof-of-concept of the approach. An important thing to keep in mind is that the amount of improvement in different applications naturally depends on the knowledge experts of that domain have, and their willingness to give the feedback. While our work dwells on improving prediction by using expert knowledge and facilitating elicitation, appropriate interface and visualization techniques are also required for a complete and effective interactive elicitation. These considerations are left for future work.

As for the applications where our method can be employed, of particular interest for future work are the medical settings, where because of the implied risks it is not always possible to increase the sample size. On the other hand, an expert may give feedback on the relevance of certain variables, or an expert might know how much some clinical and behavioural variables explain of the risk. For instance, we plan to test our methods in genomic cancer medicine cases, where the feature size is in the order of thousands (typically including gene expression, mutation data, copy number variation, and cytogenetic marker measurements), while the sample size (number of patients with known measurements) is in the order of hundreds. For different types of cancer, features that are indicative of the drug response (i.e., biomarkers) are well known in the literature (see for instance Garnett et al. 2012). Thus, based on the updated domain knowledge and experience, experts can identify and provide feedback on the relevance of some features.

Acknowledgements This work was financially supported by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN; Grants 295503, 294238, 292334, and 284642), Re:Know funded by TEKES, and MindSee (FP7ICT; Grant Agreement No. 611570). We acknowledge the computational resources provided by the Aalto Science-IT Project. We thank Juho Piironen for comments that improved the article.

Appendix A: Posterior approximation

The posterior distribution of the model and its approximation are

$$p(\boldsymbol{w}, \sigma^{-2}, \boldsymbol{\gamma} \mid \mathcal{D}) \propto p(\boldsymbol{f}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) p(\boldsymbol{f}_{\boldsymbol{w}} \mid \boldsymbol{w}) p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\sigma^{-2}) p(\boldsymbol{w} \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \approx q(\boldsymbol{w}, \sigma^{-2}, \boldsymbol{\gamma}),$$

where $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{f}_{\boldsymbol{\gamma}}, \boldsymbol{f}_{\boldsymbol{w}})$ are the training data observations together with the sets of observed expert feedback. The individual terms are approximated as

$$
\begin{aligned}
p(\boldsymbol{f}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) &= \prod_{j \in \mathcal{F}_\gamma} \left[\gamma_j \,\mathrm{Bernoulli}(f_{\gamma,j} \mid \pi) + (1-\gamma_j)\,\mathrm{Bernoulli}(f_{\gamma,j} \mid 1-\pi)\right] \\
&\approx \prod_{j \in \mathcal{F}_\gamma} \tilde{t}_{\mathrm{Bernoulli}}(\gamma_j \mid \tilde{\rho}_j^{f_\gamma}), \\
p(\boldsymbol{f}_{\boldsymbol{w}} \mid \boldsymbol{w}) &= \prod_{j \in \mathcal{F}_w} \mathrm{N}(f_{w,j} \mid w_j, \omega^2) = \prod_{j \in \mathcal{F}_w} \tilde{t}_{\mathrm{N}}(w_j \mid \tilde{\mu}_j^{f_w}, \tilde{\tau}_j^{f_w}), \\
p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) &= \mathrm{N}(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{w}, \sigma^2 \mathbf{I}) \approx \tilde{t}_{\mathrm{N}}(\boldsymbol{w} \mid \tilde{\boldsymbol{\mu}}^y, \tilde{\boldsymbol{\Gamma}}^y)\tilde{t}_{\mathrm{Gamma}}(\sigma^{-2} \mid \tilde{\alpha}^y, \tilde{\beta}^y), \\
p(\sigma^{-2}) &= \mathrm{Gamma}(\sigma^{-2} \mid \alpha_\sigma, \beta_\sigma) \\
&= \tilde{t}_{\mathrm{Gamma}}(\sigma^{-2} \mid \alpha_\sigma - 1, -\beta_\sigma), \\
p(\boldsymbol{w} \mid \boldsymbol{\gamma}) &= \prod_j \left[\gamma_j \,\mathrm{N}(w_j \mid 0, \psi^2) + (1-\gamma_j)\delta_0(w_j)\right] \\
&\approx \prod_{j \in \mathcal{F}_\gamma} \tilde{t}_{\mathrm{N}}(\gamma_j \mid \tilde{\mu}_j^w, \tilde{\tau}_j^w)\tilde{t}_{\mathrm{Bernoulli}}(\gamma_j \mid \tilde{\rho}_j^w), \\
p(\boldsymbol{\gamma}) &= \prod_j \mathrm{Bernoulli}(\gamma_j \mid \rho) = \prod_j \tilde{t}_{\mathrm{Bernoulli}}(\gamma_j \mid \mathrm{logit}(\rho)).
\end{aligned}
$$

Here, $\mathcal{F}_\gamma$ and $\mathcal{F}_w$ denote the sets of indices of the features that have received relevance feedback and weight feedback, respectively. π, ω^2, α_σ, β_σ, and ψ^2 are assumed fixed hyperparameters. $\tilde{t}_\cdot$ denote the exponential family forms of the corresponding distributions parametrized by the precision-adjusted mean and precision for normal distribution, and the natural parameters for Bernoulli and gamma distributions. Note that the terms $p(\sigma^{-2})$, $p(\boldsymbol{f}_{\boldsymbol{w}} \mid \boldsymbol{w})$, and $p(\boldsymbol{\gamma})$ need not be approximated as they are already of the correct exponential family form. The full posterior approximation follows as $q(\boldsymbol{w}, \sigma^{-2}, \boldsymbol{\gamma}) = q(\boldsymbol{w})q(\sigma^{-2})q(\boldsymbol{\gamma})$ with

$$
\begin{aligned}
q(\boldsymbol{w}) &= \mathrm{N}(\boldsymbol{w} \mid \bar{\boldsymbol{m}}, \bar{\boldsymbol{\Sigma}}), \\
q(\sigma^{-2}) &= \mathrm{Gamma}(\sigma^{-2} \mid \bar{\alpha}_\sigma, \bar{\beta}_\sigma), \\
q(\boldsymbol{\gamma}) &= \prod_j \mathrm{Bernoulli}(\gamma_j \mid \bar{\rho}_j),
\end{aligned}
$$

where the parameters can be identified from the products of the corresponding site term approximations and are

$$
\begin{aligned}
\bar{\boldsymbol{m}} &= \bar{\boldsymbol{\Sigma}}(\tilde{\boldsymbol{\mu}}^y + \tilde{\boldsymbol{\mu}}^w + \tilde{\boldsymbol{\mu}}^{f_w}), \\
\bar{\boldsymbol{\Sigma}} &= (\tilde{\boldsymbol{\Gamma}}^y + \mathrm{diag}(\tilde{\boldsymbol{\tau}}^w) + \mathrm{diag}(\tilde{\boldsymbol{\tau}}^{f_w}))^{-1}, \\
\bar{\alpha}_\sigma &= \alpha_\sigma + \tilde{\alpha}^y, \\
\bar{\beta}_\sigma &= \beta_\sigma - \tilde{\beta}^y, \\
\bar{\rho}_j &= \frac{1}{1 + \exp(-(\tilde{\rho}_j^w + \mathrm{logit}(\rho) + \tilde{\rho}_j^{f_\gamma}))},
\end{aligned}
$$

where $\mathrm{diag}(\cdot)$ is a diagonal matrix with the parameter as the diagonal and feedback term approximation parameters are zero for feedbacks that have not been observed.

Expectation propagation (EP) and variational Bayes (VB) inference are used to find the parameters of the posterior approximation (Minka 2001, 2005; Bishop 2006). Expectation

propagation for linear regression with spike and slab prior has been introduced by Hernández-Lobato et al. (2008) (see Hernández-Lobato et al. (2015) for a more extensive treatment). We update the $\tilde{t}_{\mathrm{N}}(\boldsymbol{w} \mid \tilde{\boldsymbol{\mu}}^y, \tilde{\boldsymbol{\Gamma}}^y)$ and $\tilde{t}_{\mathrm{Gamma}}(\sigma^{-2} \mid \tilde{\alpha}^y, \tilde{\beta}^y)$ term approximations using VB and all other terms using EP. The parameter update steps in the algorithm, to be iterated until convergence, are

1. $p(\boldsymbol{w} \mid \boldsymbol{\gamma})$ approximation using parallel EP update.
2. $p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2)$ approximation using VB update.
3. $p(\boldsymbol{f}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma})$ approximation using parallel EP update.

All of the computations have closed form solutions. The VB update is used, because there is no closed form EP update for the term. Importantly, a full covariance matrix in the posterior approximation of the regression weights $\boldsymbol{w}$ is retained. Alternatively, an approximate EP update, following Hernandez-Lobato et al. (2015), would be possible.

Appendix B: Bayesian experimental design

The task is to find the feedback that maximises the expected information gain:

$$j^* = \arg\max_{j \notin \mathcal{F}} \mathbb{E}_{p(\tilde{f}_j \mid \mathcal{D})} \left[\sum_i \mathrm{KL}[p(\tilde{y} \mid \boldsymbol{x}_i, \mathcal{D}, \tilde{f}_j) \parallel p(\tilde{y} \mid \boldsymbol{x}_i, \mathcal{D})] \right],$$

where $\mathcal{F}$ is the set of feedbacks that have already been given (to simplify notation, those are here assumed included in $\mathcal{D}$) and the summation over i goes over the training dataset. The evaluation of the expected information gain is described in the following.

The posterior predictive distribution is approximated as Gaussian:

$$p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathcal{D}) \approx \mathrm{N}(\tilde{y} \mid \tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{m}}, \tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{\Sigma}} \tilde{\boldsymbol{x}} + \bar{s}^2),$$

where $\bar{s}^2 = \frac{\bar{\beta}_\sigma}{\bar{\alpha}_\sigma}$ is the posterior mean approximation for the residual variance. Similarly, the posterior predictive distributions of the feedbacks for the two feedback types follow as approximate Gaussian and Bernoulli distributions:

$$\begin{aligned} p(\tilde{f}_{w,j} \mid \mathcal{D}) &\approx \mathrm{N}(\tilde{f}_{w,j} \mid \bar{m}_j, \bar{\Sigma}_{jj} + \omega^2), \\ p(\tilde{f}_{\gamma,j} \mid \mathcal{D}) &\approx \mathrm{Bernoulli}(\tilde{f}_{\gamma,j} \mid \pi \bar{\rho}_j + (1-\pi)(1-\bar{\rho}_j)). \end{aligned}$$

The information gain between the predictive distributions is

$$\begin{aligned} &\mathrm{KL}[p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathcal{D}, \tilde{f}_j) \parallel p(\tilde{y} \mid \tilde{\boldsymbol{x}}, \mathcal{D})] \\ &= \frac{1}{2} \left[\log \frac{\tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{\Sigma}} \tilde{\boldsymbol{x}} + \bar{s}^2}{\tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{\Sigma}}_{\tilde{f}} \tilde{\boldsymbol{x}} + \bar{s}^2_{\tilde{f}}} + \frac{\tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{\Sigma}}_{\tilde{f}} \tilde{\boldsymbol{x}} + \bar{s}^2_{\tilde{f}} + (\tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{m}}_{\tilde{f}} - \tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{m}})^2}{\tilde{\boldsymbol{x}}^\top \bar{\boldsymbol{\Sigma}} \tilde{\boldsymbol{x}} + \bar{s}^2} - 1 \right]. \end{aligned}$$

As running the EP algorithm to full convergence would be too costly for evaluating a large number of candidates, we approximate the posterior distribution with the new feedback with partial EP updates. This is similar to the approach of Seeger (2008) and Hernández-Lobato et al. (2013) for experimental design for sparse linear model. We consider the two types of feedback separately.

In the case of feedback directly on the regression weight, we add the corresponding site term (which is already of Gaussian form and does not need approximation, as noted above)

and do not update the approximations of the other site terms (including assuming $\bar{s}^2_{\tilde{f}} = \bar{s}^2$). The new posterior approximation of $\boldsymbol{w}$ with these assumptions is

$$\begin{aligned}\bar{\boldsymbol{\Sigma}}_{\tilde{f}_{w,j}} &= (\bar{\boldsymbol{\Sigma}}^{-1} + T\boldsymbol{e}\boldsymbol{e}^\top)^{-1},\\ \bar{\boldsymbol{m}}_{\tilde{f}_{w,j}} &= \bar{\boldsymbol{\Sigma}}_{\tilde{f}_{w,j}}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{m}} + h\boldsymbol{e}),\end{aligned} \tag{3}$$

where $\boldsymbol{e}$ is a vector of zeros except for 1 at jth element, $T = \frac{1}{\omega^2}$, and $h = \frac{\tilde{f}_{w,j}}{\omega^2}$. Notably, $\bar{\boldsymbol{\Sigma}}^{-1}$ and $\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{m}}$ are the precision and the precision-adjusted mean of the posterior approximation without the new feedback and are directly available from the previous EP approximation. The new posterior covariance is independent of the value of the feedback $\tilde{f}_{w,j}$ and it can be efficiently evaluated using the matrix inversion lemma as $\bar{\boldsymbol{\Sigma}}_{\tilde{f}} = \bar{\boldsymbol{\Sigma}} - \frac{1}{T^{-1}+\bar{\Sigma}_{jj}}\bar{\boldsymbol{\Sigma}}\boldsymbol{e}\boldsymbol{e}^\top\bar{\boldsymbol{\Sigma}}$. Furthermore, the expectation over the feedback in the expected information gain affects only the term with the squared difference of the means. This is

$$\begin{aligned}\mathbb{E}_{p(\tilde{f}_j|\mathcal{D})}\left[(\tilde{\boldsymbol{x}}^\top\bar{\boldsymbol{m}}_{\tilde{f}} - \tilde{\boldsymbol{x}}^\top\bar{\boldsymbol{m}})^2\right] &= \mathbb{E}_{p(\tilde{f}_j|\mathcal{D})}\left[\left(\frac{T_{jj}}{1+T\bar{\Sigma}_{jj}}\tilde{\boldsymbol{x}}^\top\bar{\boldsymbol{\Sigma}}\boldsymbol{e}\right)^2\left(\frac{h}{T} - \bar{m}_j\right)^2\right]\\ &= \left(\frac{T}{1+T\bar{\Sigma}_{jj}}\tilde{\boldsymbol{x}}^\top\bar{\boldsymbol{\Sigma}}\boldsymbol{e}\right)^2(\bar{\Sigma}_{jj}+\omega^2),\end{aligned}$$

where the first equality follows from substituting the Eq. 3 and using the matrix inversion lemma, and the second equality from $\frac{h}{T} = \tilde{f}_{w,j}$ and the remaining expectation being equal to the variance of the predictive distribution of the feedback.

In the case of relevance feedback, we add the corresponding site term for the feedback and run single EP update on it and the corresponding prior term $p(w_j \mid \gamma_j)$. These updates are purely scalar operations and do not require any costly matrix operations. Other site term approximations are not updated. The new posterior approximation of $\boldsymbol{w}$ with these assumptions is

$$\begin{aligned}\bar{\boldsymbol{\Sigma}}_{\tilde{f}_{\gamma,j}} &= (\bar{\boldsymbol{\Sigma}}^{-1} + T\boldsymbol{e}\boldsymbol{e}^\top)^{-1},\\ \bar{\boldsymbol{m}}_{\tilde{f}_{\gamma,j}} &= \bar{\boldsymbol{\Sigma}}_{\tilde{f}_{\gamma,j}}(\bar{\boldsymbol{\Sigma}}^{-1}\bar{\boldsymbol{m}} + h\boldsymbol{e}),\end{aligned}$$

where $T = [\bar{\boldsymbol{\Sigma}}^{-1}_{\tilde{f}_{\gamma,j}}]_{jj} - [\bar{\boldsymbol{\Sigma}}^{-1}]_{jj}$ and $h = [\bar{\boldsymbol{\Sigma}}^{-1}_{\tilde{f}_{\gamma,j}}\boldsymbol{m}_{\tilde{f}_{\gamma,j}}]_j - [\bar{\boldsymbol{\Sigma}}^{-1}\bar{m}]_j$. That is, now T and h are the changes in the precision and the precision adjusted mean in the jth feature and these are available with cheap scalar operations. The expectation over the value of the feedback in the expected information gain is in this case a sum of two terms and we evaluate both of the terms separately using the above scheme. Again, we use the matrix inversion lemma to avoid full inversions in computing the new posterior covariance.

References

Afrabandpey, H., Peltola, T., & Kaski, S. (2016). Interactive prior elicitation of feature similarities for small sample size prediction. In *Proceedings of the 25th conference on user modelling, adaptation and personalization (UMAP2017)* (to appear). arXiv preprint arXiv:1612.02802.

Amershi, S. (2012). *Designing for effective end-user interaction with machine learning*. PhD thesis, University of Washington.

Azari Soufiani, H., Parkes, D. C., & Xia, L. (2013). Preference elicitation for general random utility models. In *Uncertainty in artificial intelligence: Proceedings of the 29th conference* (pp. 596–605). AUAI Press.

Balcan, M. F., & Blum, A. (2008). Clustering with interactive feedback. In *Proceedings of the 19th international conference on algorithmic learning theory* (pp. 316–328).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics (ACL)* (pp. 187–205).

Busby, D. (2009). Hierarchical adaptive experimental design for Gaussian process emulators. *Reliability Engineering & System Safety*, *94*(7), 1183–1193.

Cano, A., Masegosa, A. R., & Moral, S. (2011). A method for integrating expert knowledge when learning Bayesian networks from data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *41*(5), 1382–1394.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304.

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(3), 411–436.

Donoho, D., & Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A*, *367*, 4273–4293.

Druck, G., Settles, B., & McCallum, A. (2009). Active learning by labeling features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 81–90).

Fails, J. A., & Olsen Jr., D. R. (2003). Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI)* (pp. 39–45).

Ferreira, G. S., & Gamerman, D. (2015). Optimal design in geostatistics under preferential sampling. *Bayesian Analysis*, *10*(3), 711–735. doi:10.1214/15-BA944.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, *483*(7391), 570–575.

Garthwaite, P. H., & Dickey, J. M. (1988). Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society Series B (Methodological)*, *50*, 462–474.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton: Chapman & Hall/CRC.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Hernández-Lobato, D., Hernández-Lobato, J. M., & Dupont, P. (2013). Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, *14*(1), 1891–1945.

Hernandez-Lobato, D., Hernandez-Lobato, J. M., & Ghahramani, Z. (2015). A probabilistic model for dirty multi-task feature selection. In F. Bach, D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning, PMLR, Lille, France, proceedings of machine learning research* (Vol. 37, pp. 1073–1082).

Hernández-Lobato, J. M., Dijkstra, T., & Heskes, T. (2008). Regulator discovery from gene expression time series of malaria parasites: A hierarchical approach. In *Advances in neural information processing systems 20 (NIPS)* (pp 649–656).

Hernández-Lobato, J. M., Hernández-Lobato, D., & Suárez, A. (2015). Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning*, *99*(3), 437–487.

Hickey, A. M., & Davis, A. M. (2003). Requirements elicitation and elicitation technique selection: A model for two knowledge-intensive software development processes. In *Proceedings of the 36th annual Hawaii international conference on system sciences (HICSS'03)—Track 3* (Vol. 3).

House, L., Scotland, L., & Han, C. (2015). Bayesian visual analytics: Bava. *Statistical Analysis and Data Mining*, *8*(1), 1–13.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., & Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, *75*(372), 845–854.

Kiefer, J., & Wolfowitz, J. (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics*, *30*(2), 271–294. doi:10.1214/aoms/1177706252.

Lu, Z., & Leen, T. K. (2007). Semi-supervised clustering with pairwise constraints: A discriminative approach. In *Proceedings of the eleventh international conference on artificial intelligence and statistics (AISTATS)* (pp. 299–306).

Martino, L., Vicent, J., & Camps-Valls, G. (2017). Automatic emulator and optimized look-up table generation for radiative transfer models. In *Proceedings of IEEE international geoscience and remote sensing symposium (IGARSS)*.

Micallef, L., Sundin, I., Marttinen, P., Ammad-ud-din, M., Peltola, T., Soare, M., Jacucci, G., & Kaski, S. (2017). Interactive elicitation of knowledge on feature relevance improves predictions in small data sets. In *Proceedings of the 22nd international conference on intelligent user interfaces (IUI'17)*.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence (UAI)* (pp. 362–369).

Minka, T. P. (2005). *Divergence measures and message passing*. Tech. rep., Microsoft Research.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., et al. (2006). *Uncertain judgements. Eliciting experts' probabilisties*. Chichester: Wiley.

Peltola, T., Marttinen, P., & Vehtari, A. (2012). Finite adaptation and multistep moves in the Metropolis–Hastings algorithm for variable selection in genome-wide association analysis. *PloS One*, *7*(11), e49,445.

Raghavan, H., Madani, O., & Jones, R. (2006). Active learning with feedback on features and instances. *Journal of Machine Learning Research*, *7*(Aug), 1655–1686.

Ravi, S. N., Ithapu, V. K., Johnson, S. C., & Singh, V. (2016). Experimental design on a budget for sparse linear models and applications. In *Proceedings of the 33nd international conference on machine learning (ICML)* (pp. 583–592).

Robert, S., Büttner, S., Röcker, C., & Holzinger, A. (2016). Reasoning under uncertainty: Towards collaborative interactive machine learning. In A. Holzinger (Ed.), *Machine learning for health informatics* (pp. 357–376). Berlin: Springer.

Ruotsalo, T., Jacucci, G., Myllymäki, P., & Kaski, S. (2014). Interactive intent modeling: Information discovery beyond search. *Communications of the ACM*, *58*(1), 86–92.

Schäfer, C., & Chopin, N. (2013). Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing*, *23*, 163–184. doi:10.1007/s11222-011-9299-z.

Seeger, M. W. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, *9*, 759–813.

Settles, B. (2010). *Active learning literature survey*. Computer Sciences technical report 1648, University of Wisconsin, Madison.

Settles, B. (2011). Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1467–1478).

Soare, M., Ammad-ud-din, M., & Kaski, S. (2016). Regression with n $\rightarrow$ 1 by expert knowledge elicitation. In *Proceedings of the 15th IEEE ICMLA international conference on machine learning and applications* (pp. 734–739).

Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, I. H. (2001). Interactive machine learning: Letting users build classifiers. *International Journal of Human-Computer Studies*, *55*(3), 281–292.

Publication III

Iiris Sundin*, Tomi Peltola*, Luana Micallef, Homayun Afrabandpey, Marta Soare, Muntasir Mamun Majumder, Pedram Daee, Chen He, Baris Serim, Aki Havulinna, Caroline Heckman, Giulio Jacucci, Pekka Marttinen, and Samuel Kaski. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34, 13, i395–i403, 2018.

Bioinformatics, 34, 2018, i395–i403
doi: 10.1093/bioinformatics/bty257
ISMB 2018

Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge

Iiris Sundin[1,†], Tomi Peltola[1,†], Luana Micallef[1], Homayun Afrabandpey[1], Marta Soare[1,‡], Muntasir Mamun Majumder[2], Pedram Daee[1], Chen He[3], Baris Serim[3], Aki Havulinna[2,4], Caroline Heckman[2], Giulio Jacucci[3], Pekka Marttinen[1,*] and Samuel Kaski[1,*]

[1]Department of Computer Science, Helsinki Institute for Information Technology HIIT, Aalto University, Helsinki, Finland, [2]Institute for Molecular Medicine Finland FIMM, Helsinki Institute of Life Science and [3]Department of Computer Science, Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, Finland and [4]National Institute for Health and Welfare THL, Helsinki, Finland

*To whom correspondence should be addressed.
[†]The authors wish it to be known that these authors contributed equally.
[‡]Present address: Université d'Orléans, INSA Centre Val de Loire, LIFO EA 4022 Orléans, France

Abstract

Motivation: Precision medicine requires the ability to predict the efficacies of different treatments for a given individual using high-dimensional genomic measurements. However, identifying predictive features remains a challenge when the sample size is small. Incorporating expert knowledge offers a promising approach to improve predictions, but collecting such knowledge is laborious if the number of candidate features is very large.
Results: We introduce a probabilistic framework to incorporate expert feedback about the impact of genomic measurements on the outcome of interest and present a novel approach to collect the feedback efficiently, based on Bayesian experimental design. The new approach outperformed other recent alternatives in two medical applications: prediction of metabolic traits and prediction of sensitivity of cancer cells to different drugs, both using genomic features as predictors. Furthermore, the intelligent approach to collect feedback reduced the workload of the expert to approximately 11%, compared to a baseline approach.
Availability and implementation: Source code implementing the introduced computational methods is freely available at https://github.com/AaltoPML/knowledge-elicitation-for-precision-medicine.
Contact: first.last@aalto.fi
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

An urgent challenge in computational biology is how to bring machine learning and statistical models closer to clinical practitioners. Toward resolving this, we study human-in-the-loop prediction, in which a medical expert interacts with a machine learning model with the goal to improve predictions for genomics-based precision medicine. In precision medicine, large-scale screening and sequencing produce thousands of genomic and molecular features for each individual, which can then be used for predicting a phenotype of interest, such as quantitative drug sensitivity scores (DSS) of cancer cells. What makes the task particularly difficult is that the sample sizes may be extremely small, possibly dozens of individuals only, or even fewer, for example, in the case of rare cancers. Statistical methods exist for learning predictive features and models in omics-based data analysis tasks and are in principle applicable across similar tasks. Commonly applied methods include multivariate analysis of variance (Garnett *et al.*, 2012) and sparse regression models, such as lasso and elastic net (Garnett *et al.*, 2012; Jang *et al.*, 2014). Kernel methods enable finding more complex nonlinear combinations of the features (Ammad-ud din *et al.*, 2016; Costello *et al.*, 2014). However, the scarcity of data poses a serious challenge for accurate prediction with any of these techniques.

One solution to the problem of small sample size is to measure more data, using, for example, active learning to design next clinical trials (Deng *et al.*, 2011; Minsker *et al.*, 2016). This, however, is often not viable due to costs, risks or the rarity of the disease. Statistical means to alleviate the problem include multitask learning to share strength between related outputs (Ammad-ud din *et al.*, 2016; Yuan *et al.*, 2016), and the use of biological prior knowledge available in data bases. For instance, knowledge about cancer pathways has been used as side information for prediction (Ammad-ud din *et al.*, 2016; Costello *et al.*, 2014), for feature selection (De Niz *et al.*, 2016; Jang *et al.*, 2015) or to modify regularization of a model (Sokolov *et al.*, 2016). Another method, complementary to these methods, is to collect prior knowledge directly from an expert. Such prior elicitation techniques (O'Hagan *et al.*, 2006) have been used for constructing prior distributions for Bayesian data analysis that take into account expert knowledge and hence can restrict the range of parameters in predictive models (Afrabandpey *et al.*, 2017; Garthwaite *et al.*, 2013; Garthwaite and Dickey, 1988; Kadane *et al.*, 1980).

The field of precision medicine poses a major challenge for eliciting prior knowledge directly from medical experts, namely the huge number of possible genomic features that the expert needs to provide feedback on. Consequently, in practice elicitation is only possible if the effort required from the expert can be minimized. The key insight in this paper is that interactive and sequential learning can help by carefully deciding what to ask from the expert. It has earlier been used in different types of tasks, for clustering (Balcan and Blum, 2008; Lu and Leen, 2007), Bayesian network learning (Cano *et al.*, 2011) and visualization (House *et al.*, 2015). We have applied it recently also to prediction using linear regression in our preliminary work (Daee *et al.*, 2017; Micallef *et al.*, 2017; Soare *et al.*, 2016). However, these methods are not immediately applicable to precision medicine due to many open questions, in particular (1) how to most effectively personalize predictions for a specific patient, (2) which of the different ways of collecting feedback interactively are the most efficient, (3) what kind of feedback most efficiently improves prediction accuracy and (4) how to handle the multi-task problem arising in multi-output settings.

In this paper, we carefully address these challenges in the context of prediction of multivariate quantitative traits from genomic features. In particular, we (i) introduce a new targeted sequential expert knowledge elicitation approach, (ii) compare it to non-targeted and baseline sequential elicitation methods, (iii) introduce and compare two kinds of feedback for precision medicine tasks and (iv) formulate and evaluate the approaches in multivariate precision medicine tasks with real medical datasets. In order to do this, we introduce a joint probabilistic model for the prediction and for the expert feedback; in detail, we use a sparse linear regression model that extends the textual-data model of Daee *et al.* (2017). The expert feedback is here extended to include information about the direction of a putative effect, in addition to indicating whether or not a particular effect is at all relevant in a given prediction problem. We then formulate two sequential methods for collecting expert feedback in the precision medicine task. The first targets improving personalized predictions for a single individual, while the second averages predictions over all individuals. Both aim at minimizing the effort required from the expert (Fig. 1).

Our main methodological innovation, in addition to the important technical extensions of including directional feedback and tailoring the sequential elicitation to the multi-task precision medicine problem, is in introducing a new targeted or personalized sequential knowledge elicitation approach, where the queries to the expert are chosen to be the most informative for predicting the phenotype of a new, previously unseen patient. The methods are evaluated empirically in this paper; our main experimental contribution is assessing the feasibility of expert knowledge elicitation for precision medicine. Our experiments consist of two parts. First, we apply the proposed methods in a realistic simulated expert setting. In particular, we show that simulated expert feedback based on a published meta-analytic genome-wide association study improves prediction of metabolite concentrations from single nucleotide polymorphisms (SNPs) and that the sequential elicitation can reap the benefit with a small number of queries to the expert. Second, and more importantly, we demonstrate the clinical potential of the proposed approach in the difficult task of predicting drug sensitivity of *ex vivo* blood cancer cells from patients, with feedback from domain experts.

2 Models and algorithms

In this section, we describe the proposed models and algorithms for sequential expert knowledge elicitation. First, we describe a sparse linear regression model that is used to learn the relationship between the features (here, genomic features) and the multivariate quantitative traits (metabolite concentrations or drug sensitivities) and which takes into account the elicited expert knowledge. Then we introduce the two elicitation methods developed for prediction tasks in precision medicine.

2.1 Prediction model

2.1.1 Sparse Bayesian linear regression

Sparse linear regression is used to predict the quantitative traits based on the genomic features. Let $y_{n,d}$ be the value of the dth trait for nth patient, and $\boldsymbol{x}_n \in \mathbb{R}^M$ be the vector of the individual's M genomic features. We assume that the trait depends linearly on the genomic features:

$$y_{n,d} \sim \mathrm{N}(\boldsymbol{w}_d^\top \boldsymbol{x}_n, \sigma_d^2),$$

where the $\boldsymbol{w}_d \in \mathbb{R}^M$ are the regression weights and σ_d^2 is the residual variance. In practice only a small number of features are expected to have any effect on the trait, and we encode this assumption using a sparsity-inducing spike-and-slab prior (George and McCulloch, 1993; Mitchell and Beauchamp, 1988) on the weights:

$$w_{d,m} \sim \gamma_{d,m}\ \mathrm{N}(0, \tau_{d,m}^2) + (1 - \gamma_{d,m})\delta_0,$$

where $\gamma_{d,m}$ is a binary variable indicating whether the mth feature is relevant (i.e. $w_{d,m}$ drawn from a zero-mean Gaussian prior with variance $\tau_{d,m}^2$) or not ($w_{d,m}$ is set to zero via the Dirac delta spike δ_0) when predicting for the dth trait. The prior probability of relevance ρ_d controls the expected sparsity of the model via the prior

$$\gamma_{d,m} \sim \mathrm{Bernoulli}(\rho_d).$$

The model is completed with the hyperpriors

$$\sigma_d^{-2} \sim \mathrm{Gamma}(\alpha_\sigma, \beta_\sigma),$$
$$\rho_d \sim \mathrm{Beta}(\alpha_\rho, \beta_\rho),$$
$$\tau_{d,m} \sim \mathrm{Log-N}(\mu, \omega^2).$$

Settings for the values of the hyperparameters are discussed within the details of the experiments (Sections 3.1.1 and 3.2.1).

Given the observed trait values $\boldsymbol{Y} \in \mathbb{R}^{N \times D}$ for N patients and D traits and the genomic features $\boldsymbol{X} \in \mathbb{R}^{N \times M}$, the posterior distribution of the model parameters $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{\tau}^2, \boldsymbol{\sigma}^2)$ is computed via the Bayes theorem as follows:

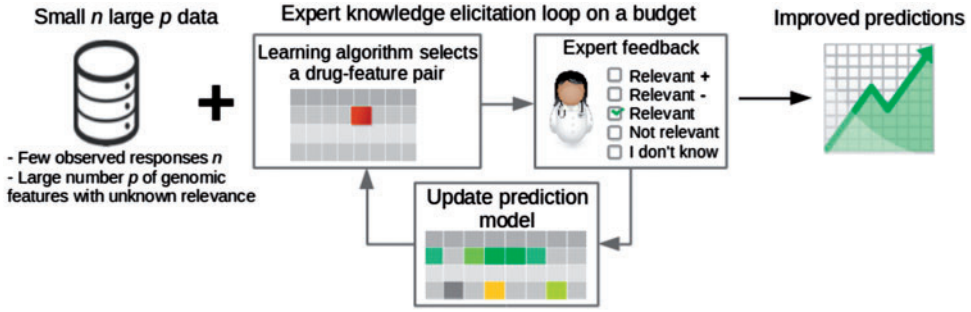

Fig. 1. Overview. Predictions in small-sample-size problems are improved by asking experts in an elicitation loop. The system presents questions for the expert sequentially to maximize performance with a minimal number of questions, i.e. on a budget. The expert answers the questions by indicating whether a feature is relevant in predicting quantitative traits, such as cancer cell's sensitivity to a drug. The expert can also indicate in which direction the effect is likely to be

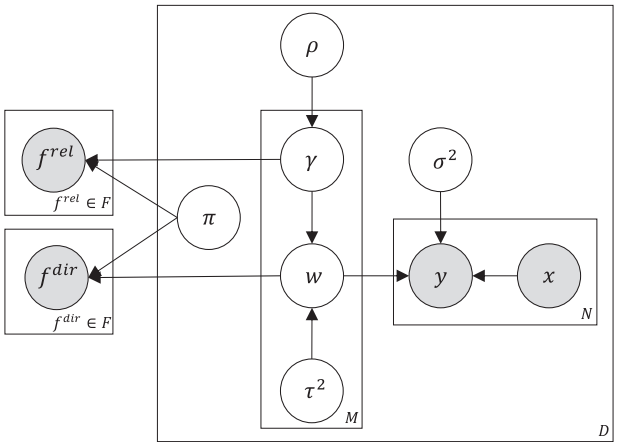

Fig. 2. Plate notation of the quantitative trait prediction model (right) and feedback observations (left) as introduced in Section 2.1. The feedbacks f^{rel} and f^{dir} are sequentially queried from the expert based on an expert knowledge elicitation method

$$p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\boldsymbol{\sigma}^2)p(\boldsymbol{w}|\gamma,\boldsymbol{\tau}^2)p(\gamma|\boldsymbol{\rho})p(\boldsymbol{\rho})p(\boldsymbol{\tau}^2)p(\boldsymbol{\sigma}^2)}{p(\boldsymbol{Y}|\boldsymbol{X})}.$$

The posterior distribution of $\boldsymbol{w}$ together with the observation model is then used to compute the predictive distribution of the traits $\tilde{y} = [\tilde{y}_1, \ldots, \tilde{y}_D]^\top$ for a new individual $\tilde{x}$:

$$p(\tilde{y}|\boldsymbol{Y},\boldsymbol{X},\tilde{x}) = \int p(\tilde{y}|\tilde{x},\boldsymbol{w},\boldsymbol{\sigma}^2)p(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X})d\boldsymbol{\theta}. \quad (1)$$

2.1.2 Incorporating expert feedback

We assume that an expert has provided feedback about the relevance of some genomic features, for example, using elicitation techniques described in the next section, corresponding to the expert's opinion of whether or not the features should be included into the model when predicting a certain trait. In addition, we assume that for some of the relevant features the expert has also indicated her expectation about the direction of the effect. These types of feedback are assumed to be available for some or all of the feature-trait pairs in the dataset, and they are treated as additional data when learning the parameters of the spike-and-slab regression model. The *relevance* feedback has been used in Daee *et al.* (2017) for univariate prediction in textual data, which we extend by including directional feedback (Micallef *et al.*, 2017) in the multi-output scenario.

Technically, the expert knowledge is incorporated into the model via feedback observation models. The relevance feedback $f^{rel}_{d,m} \in \{0,1\}$, where 0 denotes not relevant, 1 relevant, of feature m for trait d follows:

$$f^{rel}_{d,m} \sim \gamma_{d,m}\ \text{Bernoulli}(\pi^{rel}_d) + (1-\gamma_{d,m})\ \text{Bernoulli}(1-\pi^{rel}_d),$$

where π^{rel}_d is the probability of the expert being correct. For example, when the mth feature for trait d is relevant in the regression model (i.e. $\gamma_{d,m} = 1$), the expert would *a priori* be assumed to say $f^{rel}_{d,m} = 1$ with probability π^{rel}_d. In the model learning (i.e. calculating the posterior distribution in Equation (2) below), once the expert has provided the feedback based on his or her knowledge, π^{rel}_d effectively controls how strongly the model will change to reflect the feedback.

The directional feedback $f^{dir}_{d,m} \in \{0,1\}$, where 0 denotes negative weight and 1 positive, follows:

$$f^{dir}_{d,m} \sim I(w_{d,m} \geq 0)\text{Bernoulli}(\pi^{dir}_d) + I(w_{d,m} < 0)\text{Bernoulli}(1-\pi^{dir}_d),$$

where $I(C) = 1$ when the condition C holds and 0 otherwise, and π^{dir}_d is again the probability of the expert being correct. For example, when the weight $w_{d,m}$ is positive, the expert would *a priori* be assumed to say $f^{\text{dir}}_{d,m} = 1$ with probability π^{dir}_d. To simplify the model, we assume $\pi_d = \pi^{\text{dir}}_d = \pi^{\text{rel}}_d$ and set a prior on π_d as

$$\pi_d \sim \text{Beta}(\alpha_\pi, \beta_\pi).$$

Given the data $\boldsymbol{Y}$ and $\boldsymbol{X}$ and a set of observed feedbacks $\boldsymbol{F}$ encoding the expert knowledge, the posterior distribution is computed as follows:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{w},\boldsymbol{\sigma}^2)p(\boldsymbol{w}|\gamma,\boldsymbol{\tau}^2)p(\gamma|\boldsymbol{\rho})p(\boldsymbol{\rho})p(\boldsymbol{\tau}^2)p(\boldsymbol{\sigma}^2)}{p(\boldsymbol{Y},\boldsymbol{F}|\boldsymbol{X})} \times p(\boldsymbol{F}|\gamma,\boldsymbol{w},\boldsymbol{\pi})\text{p}(\boldsymbol{\pi}), \quad (2)$$

where $\mathcal{D} = (\boldsymbol{Y},\boldsymbol{X},\boldsymbol{F})$ and $\boldsymbol{\theta}$ now includes also $\boldsymbol{\pi}$. The predictive distribution follows from Equation (1). Figure 2 shows the plate diagram of the model.

The computation of the posterior distribution is analytically intractable. We use the expectation propagation algorithm (Minka and Lafferty, 2002) to compute an efficient approximation. In particular, the posterior approximation for the weights $\boldsymbol{w}$ is a multivariate Gaussian distribution and the predictive distribution for $\tilde{y}_d$ is also approximated as a Gaussian (Daee *et al.*, 2017; Hernández-Lobato *et al.*, 2015). The mean of the predictive distribution is used as the point prediction in the experimental evaluations in Section 3.

2.2 Expert knowledge elicitation methods

The purpose of expert knowledge elicitation algorithms is to sequentially select queries to the expert, such that the effort from the expert

is maximally beneficial for prediction. In univariate outcome prediction, an algorithm needs to select the next feature for an expert to provide feedback on. In the present multi-output setting, the elicitation algorithm needs to select both the output and the feature to be shown to the user in the next query. Based on preliminary experiments, we focus on sequential experimental design methods, which produced the best results for multi-output settings [Based on preliminary experiments in the multi-output setting (not shown), a Bandit model approach (Micallef *et al.*, 2017) was not better than the sequential experimental design approach by Daee *et al.* (2017)]. We next describe two new sequential experimental design methods and a baseline approach that will be compared in the results.

2.2.1 Sequential experimental design

We introduce a sequential experimental design approach to select the next (trait, feature) pair candidate, extending the work by Daee *et al.* (2017). Specifically, at each iteration t, we find the pair for which the feedback from the expert is expected to have the maximal influence on the prediction. The amount of information in the expert feedback is measured by the Kullback–Leibler divergence (KL) between the predictive distributions before and after observing the feedback. As the feedback value itself is unobserved before the actual query, an expectation over the predictive distributions of the two types of feedback is computed in finding the (trait, feature) pair (d^*, m^*) with the highest expected information gain:

$$(d^*, m^*) = \underset{(d,m)\notin \boldsymbol{F}_{t-1}}{\arg\max}\ \mathbb{E}_{\tilde{f}^{\text{rel}}_{d,m},\tilde{f}^{\text{dir}}_{d,m}|\mathcal{D}_{t-1}}\left[\sum_{n=1}^{N} u_{n,d,m,t}\right], \tag{3}$$

where $u_{n,d,m,t} = \mathrm{KL}[p(\tilde{y}_d|\boldsymbol{x}_n, \mathcal{D}_{t-1}, \tilde{f}^{\text{rel}}_{d,m}, \tilde{f}^{\text{dir}}_{d,m}) || p(\tilde{y}_d|\boldsymbol{x}_n, \mathcal{D}_{t-1})]$, $\mathcal{D}_{t-1} = (\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{F}_{t-1})$, $\boldsymbol{Y} \in \mathbb{R}^{N\times D}$ are the observed trait values for N individuals and D traits, $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ are the genomic features, and $\boldsymbol{F}_{t-1}$ is the set of feedbacks given before the current query iteration. The $u_{n,d,m,t}$ term measures the impact the feedback on feature m would have on the predictive distribution of trait d of the nth individual. The summation in n runs over the training data, and hence the criterion (3) selects the next query assuming that the individuals for whom predictions are made are similar to the training set (unlike the targeted criterion presented in the next section). Once the query (d^*, m^*) is selected and presented to the expert, the provided feedback is added to the set $\boldsymbol{F}_{t-1}$ to produce $\boldsymbol{F}_t$. Queries where the expert is not able to provide an answer do not affect the prediction model but are added to the set so as not to be repeated.

Using the approximate posterior distribution, the posterior predictive distribution of the relevance and directional feedback, $p(\tilde{f}^{\text{rel}}_{d,m}, \tilde{f}^{\text{dir}}_{d,m}|\mathcal{D}_{t-1}) = p(\tilde{f}^{\text{rel}}_{d,m}|\mathcal{D}_{t-1})p(\tilde{f}^{\text{dir}}_{d,m}|\mathcal{D}_{t-1})$, follows a product of Bernoulli distributions. The approximate posterior predictive distribution of $\tilde{y}_d$ follows a Gaussian distribution, which makes the KL divergence calculation simple. However, to make inference efficient enough for online use, we approximate the posterior with partial expectation propagation updates (Daee *et al.*, 2017; Seeger, 2008).

2.2.2 Targeted sequential experimental design

We define a new, targeted version of the sequential experimental design by computing the utility for a single new target sample instead of summing over the training dataset samples. The motivation is to try to improve the prediction specifically for the current target individual rather than overall.

For this, we maximize the following information gain:

$$(d^*, m^*) = \underset{(d,m)\notin \boldsymbol{F}_{t-1}}{\arg\max}\ \mathbb{E}_{\tilde{f}^{\text{rel}}_{d,m},\tilde{f}^{\text{dir}}_{d,m}|\mathcal{D}_{t-1}}\left[\tilde{u}_{d,m,t}\right] \text{ where}$$
$$\tilde{u}_{d,m,t} = \mathrm{KL}[p(\tilde{y}_d|\tilde{x}, \mathcal{D}_{t-1}, \tilde{f}^{\text{rel}}_{d,m}, \tilde{f}^{\text{dir}}_{d,m}) \,||\, p(\tilde{y}_d|\tilde{x}, \mathcal{D}_{t-1})],$$

where $\tilde{x}$ are the genomic features of the new, previously unseen individual. This is identical to the previous except for evaluating the information gain only at the target individual's $\tilde{x}$.

2.2.3 Random sequential sampling

As a baseline, we use uniform random sampling for the next query from the set of (trait, feature) pairs that have not yet been queried.

3 Experiments

The proposed methods are evaluated first in metabolite concentration prediction from genomic data with simulated expert feedback and then applied to real expert feedback in multiple myeloma drug sensitivity prediction. In both cases, we first compare the predictive accuracy with and without expert feedback and then assess the performance of the sequential elicitation methods.

3.1 Metabolite concentration prediction from genomic data—simulated expert feedback

We performed a simulation study of predicting the concentrations of four standard lipid profile metabolites [high-density lipoprotein cholesterol (HDL-C); low-density lipoprotein cholesterol (LDL-C); total serum cholesterol (TC); serum triglycerides (TG)] using genotype data as predictors. Both the genotypes and the metabolites were real observations, and also the feedback was simulated using real Genome-wide association study (GWAS) meta-analysis results. This setup emulates prior elicitation from a knowledgeable geneticist, who provides feedback about the relevance of different SNPs on predicting different metabolites and on the directions of the putative effects.

3.1.1 Experimental methods

The dataset comes from the Finnish FINRISK07 (DILGOM07 subset) study that sampled a random set of adults in Finland to participate in a study on general health of Finnish population (Borodulin *et al.*, 2015). We included unrelated individuals for whom genotype data and the four metabolite concentrations (measured using NMR spectroscopy) existed (Kettunen *et al.*, 2016; Marttinen *et al.*, 2014). The total number of individuals was 3918. Standard quality control was applied to the genotype data (SNP missingness rate $< 0{:}05$, minor allele frequency $> 0{:}01$, imputation quality (info) $> 0{:}3$, and HWE $> 10^{-6}$. Pairs of related individuals, as defined by pi-hat statistic $> 0{:}2$, were pruned out by removing one of them. The number of individuals after this is 3918).

We used the results of a GWAS meta-analysis of 24 925 individuals (Kettunen *et al.*, 2016) to generate the feedback and to prune the number of SNPs for consideration. The meta-analysis included the same metabolites (among others) measured using the same technology as the target metabolites here. However, the dataset we used was not included in the meta-analysis. The set of SNPs was pruned by prioritizing SNPs that had low P-values in the meta-analysis for at least one of the target metabolites and requiring that the SNPs were at least 0.125cM and 25 kb apart in the genetic map, to select a non-redundant set of SNPs. The final number of included SNPs was 3107.

Feedback was generated from the results of the meta-analysis by taking all SNPs with P-value smaller than 2.3×10^{-9} (the significance

Table 1. Performance in metabolite concentration prediction

	Data mean	Elastic net	SnS no fb	SnS all fb	SnS rel. fb
C-index	0.500	0.519	0.540	**0.558**	0.556
MSE	1.017	1.010	0.999	**0.984**	0.988
PVE	0.000	0.007	0.018	**0.032**	0.028

Note: Values are averages over the four target metabolites. Best result on each row has been boldfaced. SnS = spike and slab sparse linear model; fb = feedback; Rel. fb = Only relevance feedback; MSE = mean squared error; PVE = proportion of variance explained.

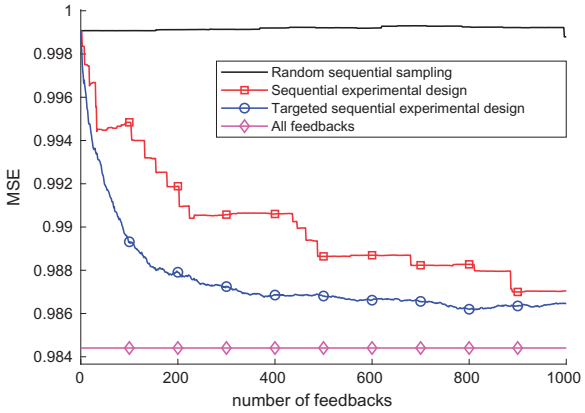

Fig. 3. Sequential experimental design performance in metabolite concentration prediction comparing random querying, information gain-based sequential experimental design and its targeted version. First 1000 iterations of feedback are shown and the result with all feedbacks is included for reference. For the targeted sequential experimental design, each individual in the test set was the target separately and the predictions in the resulting feedback sequence were used for that individual. The curve is a mean over all these sequences

threshold in the meta-analysis (Kettunen *et al.*, 2016)) as relevant (for each target separately) and those with larger than 0.9 (arbitrary; sensitivity to this is investigated in the result) as irrelevant. Directional feedback was generated for all relevant SNPs by taking the sign of the regression coefficient in the meta-analysis results. This resulted in 13, 46, 39, and 11 SNPs being considered relevant and 1010, 859 620 and 628 SNPs not relevant for HCL-C, LDL-C, TC and TG, respectively. The rest of the SNPs was considered to be of unknown relevance.

The hyperparameters of the prediction model were set as $\alpha_\sigma = 4$, $\beta_\sigma = 4$, $\alpha_\rho = 2$, $\beta_\rho = 98$, $\mu = -3.25$, $\omega^2 = \frac{1}{2}$, and $\alpha_\pi = 19$, $\beta_\pi = 1$ to reflect relatively vague information on the residual variance (roughly higher than 0.5), a preference for sparse models and small effect sizes that one expects in SNP-based regression, and the *a priori* quality of the expert knowledge as 19 correct feedbacks out of 20. A sensitivity analysis with regard to the sparsity and effect size parameters is given in the Supplementary Material.

For predictive performance evaluation, the data were divided randomly into a training set of 1000 and a test set of 2918 individuals. The proposed methods are compared against two baselines: constant prediction with the training data mean and elastic net. Elastic net is a state-of-the-art method that includes ridge and lasso regression as special cases [Elastic net is implemented using the glmnet R-package (Friedman *et al.*, 2010) with nested cross-validation for choosing the regularization parameters.]. The concordance index (C-index; the probability of predicting the correct order for a pair of samples; higher is better) (Costello *et al.*, 2014; Harrell, 2015) and the mean squared error (MSE; lower is better), computed on the test set, are used as the performance measures. Bayesian bootstrap (Rubin, 1981) over the predictions is used to evaluate the uncertainty in pairwise model comparisons: in particular, we compute the probability that model M_1 is better than model M_2 as follows Pr (M_1 is better than M_2) $= \frac{1}{B}\sum_{b=1}^{B} I(M_1$ is better than M_2 in bootstrap sample $b)$, where $I(C) = 1$ if condition C holds and 0 otherwise (Vehtari and Lampinen, 2002).

3.1.2 Simulated sequential elicitation user experiment

We simulated sequential expert knowledge elicitation by iteratively querying (metabolite, feature) pairs for feedback, and answering the queries using the generated feedback. At each iteration, the models were updated and the next query chosen, based on the feedback elicited up to that iteration, and the training data which does not change. We compared the elicitation methods described in Section 2.2.1. The queries for the targeted sequential experimental design approach were generated by running each test sample as a target individual separately. The queries were selected without replacement from the 12 428 possible queries (4 metabolites × 3, 107 SNPs).

3.1.3 Results

Expert knowledge can improve genomics-based prediction accuracy. Table 1 shows the prediction performance averaged over the four target metabolites (see Supplementary Material for target-wise performance measures; same conclusions hold for those as given here for the averaged case). As a side result, the sparse linear model without feedback (SnS no fb) improves over both baselines (data mean and elastic net), with bootstrapped model comparison probabilities for both MSE and C-index greater than 0.99 in favor of it. Next, we established whether the simulated feedback improves the model. Giving all of the feedback (SnS all fb) improves the performance (Table 1), with bootstrapped model comparison probabilities greater than 0.99 in favor of it against all other models.

Although the results show that the predictive models with feedback are confidently better, the absolute improvements in MSE are small. Yet, the amount of explanatory power in GWAS is usually small and especially when learning from small datasets. The meta-analysis results, with a much larger dataset, explained 4–11% of the variance among the four metabolites studied here (note that this is also not predictive power but computed in the same dataset as the association study). Computing the proportion of variance explained (PVE) by the cross-validated predictions, PVE $= 1 - \frac{\text{MSE}}{\text{MSE}_{\text{datamean}}}$, the improvement is 1.4 percentage points, corresponding to almost doubling (1.8×) the predictive PVE from no feedback to all feedback model (Table 1).

Feedback with the direction of the putative effect is more effective than general relevance feedback. We then examined the effect of the directional feedback compared to using relevance feedback only. Using only the relevance feedback (SnS rel. fb) improves over the no feedback model, but the performance is decreased compared to using both relevance and directional feedback (SnS all fb). We further ran a sensitivity analysis with respect to the amount of *not relevant* feedback: removing all *not relevant* feedback had a small deteriorating effect in this dataset, resulting in MSE of 0.986 and PVE of 0.031.

Sequential knowledge elicitation reduces the number of queries required from the expert. The sequential knowledge elicitation performance was then studied. Figure 3 shows the MSE as a function of the number of queried feedbacks for random, experimental design, and targeted experimental design sequential methods. The random method finds hardly any useful queries in 1000 steps. Both

experimental design methods improve over this significantly, with the targeted version being preferred overall. The targeted sequential experimental design attains 70% of the performance of the all feedback case in 122 queries (1% of all possible queries) and 80% of the performance in 257 queries (2%). This indicates that most of the benefit from the feedback can be obtained using the experimental design with much less effort from the expert than going through all the possible queries or using random selection would require.

3.2 Drug sensitivity prediction for multiple myeloma patients—real expert feedback

To evaluate the proposed methods in a realistic case, we apply them to a dataset of real patients with the blood cancer multiple myeloma and use feedback collected from two well-informed experts to simulate sequential knowledge elicitation. Details of the dataset and the expert feedback collection are presented in the next section, followed by experimental results showing the effectiveness of the methods in practice.

3.2.1 Experimental methods

We used a complete set of measurements on *ex vivo* drug sensitivities, somatic mutations and karyotype data (cytogenetic markers), generated for a cohort of 44 multiple myeloma patient samples. Drug sensitivities are presented as quantitative DSS as described by Yadav *et al.* (2014) and were calculated for 308 drugs that have been tested for dose–response in the cancer samples in five different concentrations over a 1000-fold concentration range. Somatic mutations were identified from exome sequencing data and annotated as described earlier by Kontro *et al.* (2014).

We focus our analysis on 12 targeted drugs, grouped in 4 groups based on their primary targets (BCL-2, glucocorticoid receptors, PI3K/mTOR, and MEK1/2). Also, among the mutations, we focus our analysis on those present in more than one patient. This results in data matrices of 44×12 (samples versus drugs), $44 \times 2,935$ (samples versus mutations) and 44×7 (samples versus cytogenetic markers). In this paper, we ask the experts only about the somatic mutations and cytogenetics markers, which the experts know better and hence need to spend less time on in the experiments. We will extend to molecular features with less well known effects, such as gene expression, in follow-up work.

We use leave-one-out cross-validation (That is, in computing the predictions for each patient, that particular patient is not used in learning the prediction model.) to estimate the performances of the drug sensitivity prediction models, with the C-index (the probability of predicting the correct order for a pair of samples; higher is better) (We note that C-index computed from leave-one-out cross-validation can be biased as it compares predictions for pairs of samples.We do not expect this to favor any particular method.) (Costello *et al.*, 2014; Harrell, 2015) and the MSE (lower is better) as the performance measures. MSE values are given in the normalized DSS units (zero mean, unit variance scaling on training data). Bayesian bootstrap (Rubin, 1981) over the predictions is used to evaluate the uncertainty in pairwise model comparisons (see Section 3.1.1).

The hyperparameters of the prediction model were set as $\alpha_\sigma = 4$, $\beta_\sigma = 4$, $\alpha_\rho = 1$, $\beta_\rho = 2$, $\mu = -2.5$, $\omega^2 = \frac{1}{2}$ and $\alpha_\pi = 19$, $\beta_\pi = 1$ to reflect our assumptions of relatively vague information on the residual variance (roughly higher than 0.5), a minor preference for sparse models and moderate effect sizes and the *a priori* quality of the expert knowledge as 19 correct feedbacks out of 20.

Table 2. Feedback type and count, given to the 1944 (drug, feature) pairs by the experts

Answer	SR	DC
Relevant, positive correlation	192	47
Relevant, negative correlation	14	34
Relevant, unknown correlation direction	26	358
Not relevant	13	0
I don't know	1699	1505
Total	1944	1944

Note: SR = Senior researcher, DC = Doctoral candidate.

3.2.2 Feedback collection

We collected feedback from two well-informed experts of multiple myeloma, using a form containing genes with mutations that have been causally implicated in cancer (Forbes *et al.*, 2015) (155 genes in our data), and seven cytogenetic markers, in total 162 features. The experts were asked to give feedback on the relevance of features and the direction of their effect for predicting the sensitivity to 12 targeted drugs, grouped by the targets (BCL-2, glucocorticoid receptors, PI3K/mTOR and MEK1/2). We note that the experts indicated that the same feedback applies to all drugs in the same drug group. The answer counts by feedback type are summarized in Table 2 for both of the experts. The experts were instructed not to refer to external databases while completing the feedback form, in order to collect their (tacit) prior knowledge on the problem and make the task faster for them.

3.2.3 Simulated sequential elicitation user experiment

Similar to the metabolite prediction experiment (Section 3.1.2), we simulate sequential expert knowledge elicitation by iteratively querying (drug, feature) pairs for feedback and answering the queries using the pre-collected feedback described in Section 3.2.2. The queries are selected without replacement from the 1944 pairs (12 drugs $\times$ 162 genomic features) included in the feedback collection. The rest of the mutation data (2780 mutations) are not queried for feedback, but all 2942 genomic features are included in the prediction model.

3.2.4 Results

Expert knowledge elicitation improves the accuracy of drug sensitivity prediction. Table 3 establishes the baselines by comparing the prediction model we use, the spike-and-slab regression model without expert feedback, to constant prediction of training data mean and elastic net regression (see Supplementary Material for drug-wise performance measures). Elastic net has poor performance with regard to MSE on this dataset, while the spike-and-slab model performs better.

The main result is that the complete sets of feedback from both of the experts improves the predictions, as can be seen in Table 4, which compares the spike-and-slab model without feedback to the model incorporating all available expert feedback. The model with feedback from the senior researcher has 4% higher C-index and 2% lower MSE compared to the no feedback model and is confidently better according to the bootstrapped probabilities (0.80 for C-index and 0.97 for MSE).

Feedback with the direction of the putative effect is more effective than general relevance feedback. We also assess the importance of the type of the feedback by comparing a spike-and-slab model with relevance only feedback (interpreting potential expert

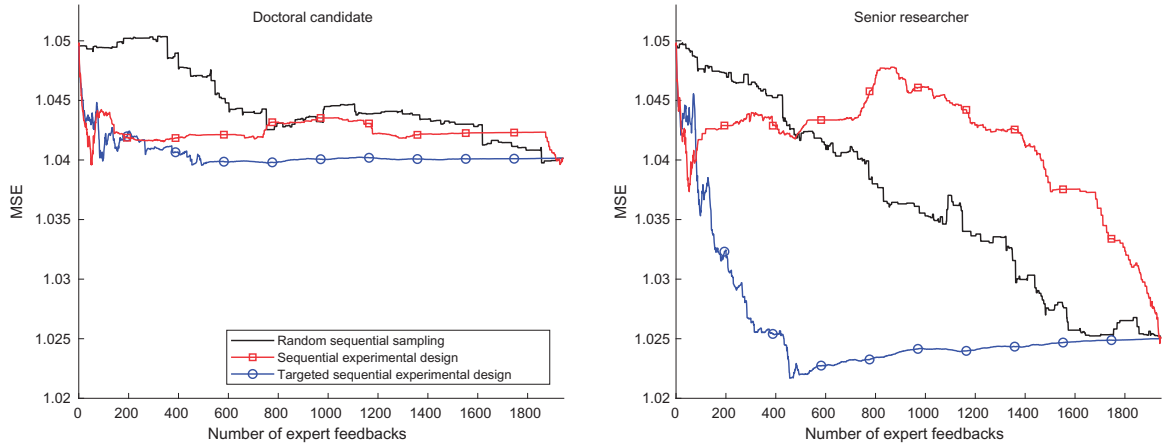

Fig. 4. Performance improves faster with the active elicitation methods than with randomly selected feedback queries. The curves show MSEs as a function of the number of iterations for the three query methods, with feedback of the doctoral candidate (left) and senior researcher (right). In each iteration, a (drug, feature) pair is queried from the expert

Table 3. Performance of drug sensitivity prediction without expert feedback

	Data mean	Elastic net	Spike-and-slab
C-index	0.500	0.505	**0.577**
MSE	1.079	1.153	**1.050**

Note: Values are averaged over the 12 drugs. Best result on each row has been boldfaced.

Table 4. Predictive performance of spike-and-slab regression with and without expert feedback

	No feedback	Doctoral candidate	Senior researcher
C-index	0.577	0.582	**0.597**
MSE	1.050	1.040	**1.025**

Note: Values are averaged over the 12 drugs.

knowledge on the direction only as relevance) to a model with both types of feedback. Table 5 shows that the directional feedback improves the performance markedly, especially in the case of the senior researcher (who gave more directional feedback than the doctoral candidate; see Table 2). The bootstrapped probabilities are 0.79 in the C-index and 0.96 in the MSE in favor of both types of feedback compared to relevance only feedback for the senior researcher and, similarity, 0.50 and 0.85 in the case of doctoral candidate. For the senior researcher, we also tested discarding all 'not-relevant' feedback (doctoral candidate didn't give any): this didn't have a noticeable effect on the performance (MSE: 1.025).

Sequential knowledge elicitation reduces the number of queries required from the expert. In the results presented so far, the experts had evaluated all (drug, feature) pairs and given their answers. We next present the main result, of how much the sequential knowledge elicitation models are able to reduce the impractical workload of the experts to give feedback on all drug-feature-pairs. We compare the effectiveness of the elicitation methods developed in this paper using a simulated user experiment (see Section 3.2.3). The results in Figure 4 show that both methods achieve faster improvement in

Table 5. Performance of drug sensitivity prediction with only relevance feedback and with relevance and directional feedback

	Doctoral candidate		Senior researcher	
	Relevance fb	All fb	Relevance fb	All fb
C-index	**0.583**	0.582	0.578	**0.597**
MSE	1.048	**1.040**	1.048	**1.025**

Note: Values are averaged over the 12 drugs.

prediction accuracy than the random selection, as a function of the amount of feedback. With sequential knowledge elicitation, 80% of the final improvement is reached in the first 230 (81) and 1871 (35) feedbacks for the targeted experimental design and non-targeted experimental design methods, respectively, using senior researcher feedback (doctoral candidate feedback). For comparison, 1362 (1619) feedbacks are required for similar accuracy if the queries are chosen randomly. Thus, on average, the targeted sequential experimental design requires only 11% (senior researcher: 17%, doctoral candidate: 5%) of the number of queries compared to random elicitation order, and the sequential experimental design model 70% [SR: 137%, DC: 2% (The improvement, however, is not stable for doctoral candidate for sequential experimental design)], to achieve 80% of the potential improvement.

4 Discussion and conclusion

Our goal was to study open questions in expert knowledge elicitation in the context of precision medicine. In summary, we introduced expert knowledge elicitation methods for and studied their feasibility in the challenging task of prediction in precision medicine. To our knowledge, this kind of approach has not been evaluated previously in precision medicine. Our results show that accumulating expert knowledge with intelligent, experimental design-based algorithms can improve the predictive performance in an efficient manner considering the effort from the expert. This is particularly important as evaluating the queries can be time-consuming for the expert, and involve searching through databases, literature and data (although here, in the real expert experiment, we evaluated the algorithms based on the tacit knowledge of two well-informed experts).

To address the individualized prediction task characteristic to precision medicine, we introduced a targeted sequential expert knowledge elicitation algorithm that sequentially selects queries that will have the greatest effect locally close to the target patient, as opposed to maximizing the effect of feedback globally over the training set of patients. In both of our experiments with real-world medical datasets, with simulated feedback and with real expert feedback, the targeted method performed clearly better than the general experimental design algorithm (and the random sampling based baseline). The developed elicitation algorithms also address the multivariate aspect of predicting for multiple quantitative traits simultaneously, which is particularly important in cases where the predictions are to be used in support of deciding, for example, between multiple alternative treatment strategies.

Our experiments showed that even relatively limited feedback may improve predictions in real-world precision medicine. In general, we expect feedback to be the most useful when the amount of data is limited, making learning of accurate effects challenging. With a lot of data, the prior distributions, and hence the feedback, are expected to have a smaller impact. Also, in extreme cases, it could happen that none of the features has any real influence on the output variable, in which case no model, with feedback or not, will be able to improve beyond the simplest mean prediction; however, such extreme situations seem unrealistic in many real-world precision medicine problems.

Furthermore, we studied the usefulness of different types of feedback. Our elicitation algorithm proceeded by selecting an input–output pair to be evaluated by an expert, and two kinds of feedback were considered: whether the genomic input feature has an effect on the output variable (relevance feedback), and, if it does, what is the direction of the putative effect (directional feedback). Our experiments indicated that including directional feedback improves upon using relevance feedback only and can often be provided without any extra effort by the expert. Nevertheless, the relevance feedback (without direction) is also needed because sometimes specifying the direction may be difficult for the expert. The directional feedback effectively halves the space of values a regression weight can take, and it can be seen as a simple case of general monotonicity constraints found useful in health care related analyses (Riihimäki and Vehtari, 2010). Of the two possible choices of relevance feedback, *relevant* or *not relevant*, we found the former much more important. It is also debatable how reliably an expert may deem some genomic feature as *not relevant*, because scientific studies rarely provide statistical evidence *against* any effect.

A natural question for future studies is how willing the experts are to use such a system. For example, if the outcome is well predicted in general, the experts may not be willing to invest time in the interaction. This potential future direction also relates to interface design, to convey the meaningfulness of the interaction to the expert. Another future direction would be to extend the model to incorporate feedback from multiple experts, which could be useful by averaging out any incorrect or biased answers a single expert might occasionally provide. Currently, our model has a parameter (π) reflecting the probability that the expert is correct, and in the extension multiple such parameters might be introduced, corresponding to experts of different levels of credibility.

The methods introduced here for precision medicine can be placed into the wider context of augmented intelligence tasks, in which a human expert works together with a machine learning system to achieve a common goal. In specific applications, some of the expert's knowledge may already be found in databases. Naturally, any reliable and structured information from databases should be built into the predictive model automatically, to save the effort from the expert. However, not all informative data are available in a structured format that could be easily incorporated and, for example, the natural language processing capabilities of machines cannot yet match the quality of human curators. Moreover, expert knowledge elicitation and incorporating data mining-based information are complementary rather than redundant. Active knowledge elicitation could, for example, be used to query an expert about the correctness or reliability of database information. Yet most importantly, the doctors and researchers will anyway be analyzing their data, even if in many cases sophisticated tools incorporating comprehensive prior knowledge will not be available in practice. In these cases not taking the experts' knowledge and expertise into account would neglect an important data source, when the lack of data may be a significant problem.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work was supported by the Academy of Finland [Finnish Center of Excellence in Computational Inference Research COIN, grant nos. 295503, 294238, 292334, 284642, 305780, 286607 and 294015], by Jenny and Antti Wihuri Foundation and by Alfred Kordelin Foundation.

Conflict of Interest: none declared.

References

Afrabandpey,H. *et al*. (2017). Interactive prior elicitation of feature similarities for small sample size prediction. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 265–269. ACM.

Ammad-Ud Din,M. *et al*. (2016) Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, **32**, i455–i463.

Balcan,M.-F. and Blum,A. (2008). Clustering with interactive feedback. In *International Conference on Algorithmic Learning Theory*, pp. 316–328. Springer.

Borodulin,K. *et al*. (2015) Forty-year trends in cardiovascular risk factors in Finland. *Eur. J. Public Health*, **25**, 539–546.

Cano,A. *et al*. (2011) A method for integrating expert knowledge when learning Bayesian networks from data. *IEEE Trans Syst Man Cybern B Cybern.*, **41**, 1382–1394.

Costello,J.C. *et al*. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, **32**, 1202–1212.

Daee,P. *et al*. (2017) Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Mach. Learn.*, **106**, 1599–1620.

De Niz,C. *et al*. (2016) Algorithms for drug sensitivity prediction. *Algorithms*, **9**, 77.

Deng,K. *et al*. (2011). Active learning for developing personalized treatment. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI'11)*, pp. 161–168.

Forbes,S.A. *et al*. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Garnett,M.J. *et al*. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Garthwaite,P.H. *et al*. (2013) Prior distribution elicitation for generalized linear and piecewise-linear models. *J. Appl. Stat.*, **40**, 59–75.

Garthwaite,P.H. and Dickey,J.M. (1988) Quantifying expert opinion in linear regression problems. *J. Roy. Stat. Soc. Ser. B (Methodological)*, **50**, 462–474.

George,E.I. and McCulloch,R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

Harrell,F. (2015) *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd edn. Springer, Cham.

Hernández-Lobato,J.M. *et al.* (2015) Expectation propagation in linear regression models with spike-and-slab priors. *Mach. Learn.*, **99**, 437–487.

House,L. *et al.* (2015) Bayesian visual analytics: baVa. *Stat. Anal. Data Mining*, **8**, 1–13.

Jang,I.S. *et al.* (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Pacific Symposium on Biocomputing*, pp. 63–74.

Jang,I.S. *et al.* (2015). Stepwise group sparse regression (SGSR): gene-set-based pharmacogenomic predictive models with stepwise selection of functional priors. In *Pacific Symposium on Biocomputing*, Vol. 20, pp. 32–43.

Kadane,J.B. *et al.* (1980) Interactive elicitation of opinion for a normal linear model. *J. Am. Stat. Assoc.*, **75**, 845–854.

Kettunen,J. *et al.* (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.*, **7**, 11122.

Kontro,M. *et al.* (2014) Novel activating STAT5B mutations as putative drivers of T-cell acute lymphoblastic leukemia. *Leukemia*, **28**, 1738–1742.

Lu,Z. and Leen,T.K. (2007). Semi-supervised clustering with pairwise constraints: a discriminative approach. In *Proc of AISTATS*, pp. 299–306.

Marttinen,P. *et al.* (2014) Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics*, **30**, 2026–2034.

Micallef,L. *et al.* (2017). Interactive elicitation of knowledge on feature relevance improves predictions in small data sets. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pp. 547–552, New York, NY, USA, ACM.

Minka,T.P. and Lafferty,J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pp. 352–359.

Minsker,S. *et al.* (2016) Active clinical trials for personalized medicine. *J. Am. Stat. Assoc.*, **111**, 875–887.

Mitchell,T.J. and Beauchamp,J.J. (1988) Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, **83**, 1023–1032.

O'Hagan,A. *et al.* (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, Chichester, England.

Riihimäki,J. and Vehtari,A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS2010*, pp. 645–652.

Rubin,D.B. (1981) The Bayesian bootstrap. *Ann. Stat.*, **9**, 130–134.

Seeger,M.W. (2008) Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, **9**, 759–813.

Soare,M. *et al.* (2016). Regression with n → 1 by expert knowledge elicitation. In *Proceedings of the 15th IEEE ICMLA International Conference on Machine learning and Applications*, pp. 734–739.

Sokolov,A. *et al.* (2016) Pathway-based genomics prediction using generalized elastic net. *PLoS Comput. Biol.*, **12**, e1004790.

Vehtari,A. and Lampinen,J. (2002) Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput.*, **14**, 2439–2468.

Yadav,B. *et al.* (2014) Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci. Rep.*, **4**, 5193.

Yuan,H. *et al.* (2016) Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.*, **6**, 31619.

Publication IV

Pedram Daee*, Tomi Peltola*, Aki Vehtari, and Samuel Kaski. User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, Tokyo, Japan, 305–310, March 2018.

User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction

Pedram Daee[†], **Tomi Peltola**[†], **Aki Vehtari, Samuel Kaski**
Helsinki Institute for Information Technology HIIT
Department of Computer Science
Aalto University, Espoo, Finland
firstname.lastname@aalto.fi
[†]Authors contributed equally.

ABSTRACT
In human-in-the-loop machine learning, the user provides information beyond that in the training data. Many algorithms and user interfaces have been designed to optimize and facilitate this human–machine interaction; however, fewer studies have addressed the potential defects the designs can cause. Effective interaction often requires exposing the user to the training data or its statistics. The design of the system is then critical, as this can lead to double use of data and overfitting, if the user reinforces noisy patterns in the data. We propose a user modelling methodology, by assuming simple rational behaviour, to correct the problem. We show, in a user study with 48 participants, that the method improves predictive performance in a sparse linear regression sentiment analysis task, where graded user knowledge on feature relevance is elicited. We believe that the key idea of inferring user knowledge with probabilistic user models has general applicability in guarding against overfitting and improving interactive machine learning.

Author Keywords
Interactive machine learning; probabilistic modelling; Bayesian inference; overfitting; expert prior elicitation; human-in-the-loop machine learning;

INTRODUCTION
Interactive and human-in-the-loop machine learning (and the related field of visual analytics) exploits the complementary knowledge and skills of humans and machines to improve performance over automatic training-data-based machine learning and to extend the reach of machine learning systems [1, 2, 3, 8, 13, 17]. However, the study of how to optimally combine the strengths of humans and machines is still in its early phase, and many possible issues arising from the interaction have not been thoroughly considered yet.

IUI'18, March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-4945-1/18/03…$15.00
DOI: https://doi.org/10.1145/3172944.3172989

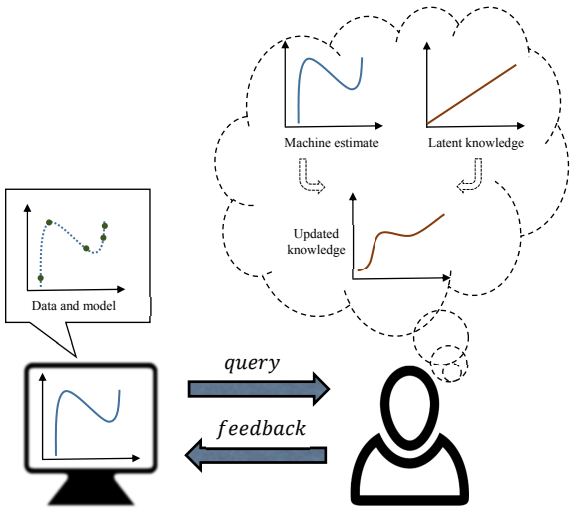

Figure 1. A schematic example of overfitting in human-in-the-loop machine learning. The machine and user are collaborating to improve a regression problem. The machine fits a regression model to the training data (green dots) and employs it to interact with the user (here by visualizing the trained model). Through the interaction, the user learns aspects of the training data and considers them to update her latent knowledge and to form her feedback. This creates a dependency between the feedback and training data that needs to be accounted for in the model to avoid double use of data and overfitting.

Overfitting, that is, the model fitting to idiosyncrasies in the training data and hence not generalizing to new data, is a thoroughly studied topic in automatic machine learning. Subtle issues with regard to overfitting can arise when introducing a human into the loop. Yet the risk of overfitting seems little discussed in the interactive machine learning literature, although many methods combine user interaction with training-data-based machine learning models. For example, many methods visualize statistics of training data or machine output directly for the user [1, 10, 11, 18, 13, 14, 21, 23] or use the training data to select informative queries to present for the user (e.g., active learning) [3, 19, 20]. Some methods use *validation* datasets in addition to the training set to evaluate the performance. This can help to alleviate overfitting, but if the user can interact with the model based on the results on the validation, the validation set is effectively only another training

dataset. A parallel line of research into controlling for overfitting has spawned in adaptive data analysis [4, 16], which has a related but different goal from interactive machine learning, of exploring the data for interesting hypotheses.

In interactive machine learning, we note that, in particular, the following three steps can lead to overfitting and hence decrease in the performance of the model, as they violate the assumption of independence of the feedback and the training data: (1) showing the training data or some of its statistics to the user, (2) querying the user for feedback, and (3) inputting the feedback back to the model as independent data (a common assumption in machine learning models). Figure 1 illustrates the induced user behaviour producing a dependency between the feedback and training data. Overfitting can also happen if the user controls some preferences (e.g., cost function weighting or regularization) in the model based on the training or validation data, since the effected improvement in the fit to the training or validation data will not necessarily generalize. The more freedom the user has (or, in this sense, the richer the feedback is), the more problematic this can be.

Rather than tying the hands of the user, improved performance can be attained by accounting for the risk of overfitting in the design of the interactive machine learning system. We propose a user modelling approach in probabilistic models to infer the user's knowledge that is complementary to the training data. We assume that the user behaves rationally, in a simple Bayesian sense [6], in combining her latent knowledge with the information that the machine reveals of the training data. Given the observed user feedback, we then invert the process to infer the latent user knowledge and use it to update the model. We illustrate the approach in a proof-of-concept user study in a sparse linear regression prediction task, where graded feature relevance knowledge is elicited from the users.

METHOD

Overview

We consider human–machine interaction in probabilistic models. In particular, we consider a situation, where the machine estimates a probabilistic model from training data, and then asks the user for feedback for the learned model. The feedback is included as further data in the model. This procedure can be iterated, although in the specific implementation here, we only consider a single round.

The setup is as follows. The machine has a probabilistic model for a prediction task and a set of training data that it uses to fit the parameters of the model. The machine assumes that the user has knowledge about some aspect of the task and elicits user knowledge by displaying the result of learning from the training data and asking for feedback. This kind of interaction is common in interactive machine learning. If the system naively uses this user feedback to update the model, for example, including it as an observation that is assumed independent of the training data, it risks double use of the data and overfitting, because such assumption could be easily violated. On the other hand, building a feedback model that would adequately describe the dependency of the feedback on the training data can be difficult.

We instead propose to infer the latent (unobserved) user knowledge, representing information that the user has beyond that of the training data, from the observed user feedback. This inferred knowledge can then be used to update the machine's model without double use of the training data. To make this feasible, we assume that the user behaves rationally, using the Bayes theorem, in integrating the information sources (her knowledge beyond training data and the information the machine provided). We then invert the process to infer the latent user knowledge.

General Mathematical Formulation

Let the observation model of a training dataset $\mathscr{D}$ be $p(\mathscr{D} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector containing the model parameters, and $p(\boldsymbol{\theta})$ be their prior distribution. Given the model and the training dataset, the machine computes the posterior distribution $p(\boldsymbol{\theta} \mid \mathscr{D})$ of the parameters using the Bayes theorem. We assume the user has knowledge about some parameter or statistic ϕ, which is an element in $\boldsymbol{\theta}$ or, more generally, a function of $\boldsymbol{\theta}$. The machine provides the user with information on ϕ, for example, its posterior distribution $p(\phi \mid \mathscr{D})$ and asks the user to provide feedback on it.

Let f be the latent (unobserved) user knowledge. Our goal is to infer from the observed feedback a latent feedback likelihood function $p(f \mid \phi)$ that can be used to update the model to $p(\boldsymbol{\theta} \mid \mathscr{D}, f)$. By the assumption of a rational user, the observed feedback is based on the posterior distribution

$$p(\phi \mid \mathscr{D}, f) = \frac{p(f \mid \phi) p(\phi \mid \mathscr{D})}{p(f \mid \mathscr{D})}, \tag{1}$$

where $p(f \mid \mathscr{D}) = \int p(f \mid \phi) p(\phi \mid \mathscr{D}) d\phi$ is the normalization constant. The technical details on how to invert this to learn $p(f \mid \phi)$ are case-specific. In the next section, we will show how to do this in eliciting feature relevance for sparse linear regression.

Feature Relevance Elicitation in Sparse Linear Regression

We apply the approach to infer user knowledge on feature relevance in sparse linear regression. We use a probabilistic sparse linear regression model described in [3], which formulates a linear model to predict the target variable y given a vector of features $\boldsymbol{x}$ and uses a spike-and-slab prior to model whether features are included or excluded from the regression:

$$\begin{aligned} y_i &\sim \mathrm{N}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}, \sigma^2), \\ \sigma^{-2} &\sim \mathrm{Gamma}(\alpha_\sigma, \beta_\sigma), \\ w_j &\sim \gamma_j \mathrm{N}(0, \tau^2) + (1 - \gamma_j) \delta_0, \\ \gamma_j &\sim \mathrm{Ber}(\rho), \end{aligned}$$

where $i = 1, \ldots, N$ runs over N training samples $(y_i, \boldsymbol{x}_i) \in \mathscr{D}$, $j = 1, \ldots, M$ runs over M features, σ^2 is a residual variance parameter, the w_j are regression weights, and the γ_j are binary variables indicating whether feature j is included in the regression ($\gamma_j = 1$: w_j is a priori normally distributed with variance τ^2) or excluded ($\gamma_j = 0$: $w_j = 0$ via the point mass δ_0). The parameter ρ is the prior expected proportion of included variables. The probabilistic parameters of the model

are $\boldsymbol{\theta} = (\boldsymbol{w}, \sigma^2, \boldsymbol{\gamma})$; here, α_σ, β_σ, τ^2, and ρ are fixed hyperparameters.

To elicit feedback on feature relevance, we show the marginal posterior probability $p(\gamma_j = 1 \mid \mathscr{D})$ to the user and ask her to provide as feedback her estimate of the probability of the feature being relevant for the prediction. To infer the latent user knowledge, we assume that the user is rational in integrating her latent knowledge f_j and the information (i.e., $p(\gamma_j = 1 \mid \mathscr{D})$) revealed of the training data. In particular, we assume that the user's observed feedback is the posterior probability

$$p(\gamma_j = 1 \mid \mathscr{D}, f_j) = \frac{p(f_j \mid \gamma_j = 1) p(\gamma_j = 1 \mid \mathscr{D})}{Z}, \quad (2)$$

where Z is the normalization constant and the latent feedback likelihood, representing the latent user knowledge, is

$$p(f_j \mid \gamma_j) = A_{f_j} \gamma_j + B_{f_j}(1 - \gamma_j),$$

where A_{f_j} is the likelihood for the latent f_j when $\gamma_j = 1$ and B_{f_j} when $\gamma_j = 0$. Without loss of generality (for using the likelihoods for updating the model later), we can set $A_{f_j} + B_{f_j} = 1$, with $A_{f_j} \in (0,1)$ and $B_{f_j} \in (0,1)$.

We infer A_{f_j} (and, consequently, $B_{f_j} = 1 - A_{f_j}$) by solving from the Bayes theorem in Equation 2:

$$A_{f_j} \propto \frac{p(\gamma_j = 1 \mid \mathscr{D}, f_j)}{p(\gamma_j = 1 \mid \mathscr{D})},$$

where the numerator is the observed feedback given by the user and the denominator is the machine's posterior probability that was shown to the user.

Given a set of observed feedbacks $\mathscr{F}$ from the user (consisting of a set of $p(\gamma_j = 1 \mid \mathscr{D}, f_j)$ values), we update the model to $p(\boldsymbol{\theta} \mid \mathscr{D}, \mathscr{F})$, where $\boldsymbol{\theta}$ denotes all parameters of the model. This is done by updating $p(\boldsymbol{\theta} \mid \mathscr{D})$ using the Bayes theorem with the inferred likelihood functions $p(f_j \mid \gamma_j)$ for each feature j with observed feedback.

Computation of the posterior distribution is intractable. Expectation propagation is used to approximate it (see [3]).

EXPERIMENT AND RESULTS

Sentiment Analysis Task

We considered the problem of rating prediction from user reviews on Amazon kitchen products. The task and data were previously studied in [3, 7]. The data consist of review texts, represented as bag-of-words with 824 distinct unigram and bigram keywords, and their corresponding 1–5 star ratings. The machine learning system aims to predict the ratings of new reviews (test data), given some training data and external expert knowledge. To make the prediction challenging, the data set was randomly partitioned in 500 training data and 4649 test data (the number of training data is smaller than the number of dimensions).

The goal of the experiment was introduced to the participants as eliciting domain knowledge from people to help in designing better predictors of product ratings. However, the real research question was to investigate the efficiency of the user model and user interaction in different scenarios. We listed 70 keywords from the reviews and asked the participants to judge the probability of relevance of each keyword in predicting the product rating. The participants could provide feedback by adjusting the probability value between 0 (not-relevant at all) to 1 (absolutely relevant) by moving a slider. The feedback was only recorded after the slider was moved. The participants could skip giving feedback to keywords that they were very uncertain about. The task description provided example keywords *must buy*, *disaster*, and *is* and explained that the first two provide useful information about the product rating, and therefore, they are relevant while the latter is uninformative for rating prediction.

	Test MSE ± STD		
Condition	*No feedback*	*User feedback*	*User model*
Baseline	1.835	1.749 ± 0.050	NA
IE	1.835	1.744 ± 0.045	1.705 ± 0.038

Table 1. Mean and standard deviation (STD) of MSE on test data for the two systems in different conditions.

Two systems were implemented, a baseline system where no information about the training data is revealed to the user and the initial positions of the sliders were set to the default value 0, and an interactive elicitation (IE) system where the initial positions were set by the machine to the posterior inclusion probabilities $p(\gamma_j = 1 \mid \mathscr{D})$ based on the training data. The participants were informed about the initialization method. Both systems use the prediction model and the feedback likelihoods introduced in the previous section with the difference that IE can infer the latent user knowledge based on the proposed user model while the baseline directly applies the user feedback in the model. Following [3], the model hyperparameters were set as $\alpha_\sigma = 1$, $\beta_\sigma = 1$, $\rho = 0.3$, and $\tau^2 = 0.01$. Mean squared error (MSE) on test data is used as the performance measure. The codes, data, and experiment forms can be found in https://github.com/HIIT/human-overfitting-in-IML.

Results

48 university students and researchers participated in the user study, 3 were excluded since they either left more than $2/3$ of the questions unanswered or they finished the study in less than 3 minutes (it was unrealistic to go through the form in less than 3 minutes). 23 participants used the baseline and the remaining 22 the IE system.

Table 1 shows the average test MSE for participants of the two systems before and after receiving feedback and also after the correction done by the proposed user model in IE. The feedback improved the predictive performance in both systems (p-value$= 3 \times 10^{-8}$ in baseline and p-value$= 5 \times 10^{-9}$ in IE without user model, using paired-sample t-test between participants before considering any feedback and after directly using the feedbacks). This shows, as analogously claimed in [3], that the participants, on average, have the necessary knowledge to improve the prediction. More prominently, the predictive performance further improved in the IE system after inferring the latent user knowledge using the user model (p-

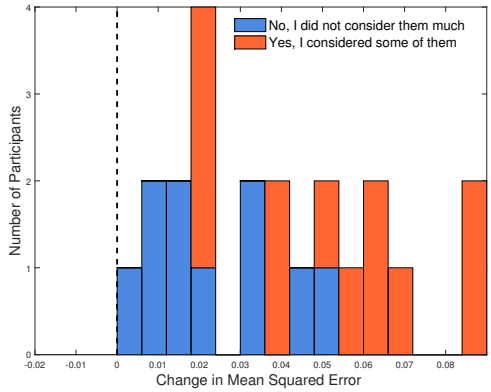

Figure 2. Stacked histogram of MSE change (the difference between MSE after directly using the feedbacks and MSE after the correction done by the propose user model) for 22 participants in IE system. The participants are grouped based on their answer to the question on whether they found the machine estimates useful or not. User modelling has improved MSE values for all participants.

value$= 7 \times 10^{-7}$ in paired-sample t-test between directly using the feedbacks and after employing the user model). Moreover, the user model improved the predictions for each individual user (Figure 2). In a post-questionnaire, we asked the participants of the IE system about the usefulness of the machine estimates: 12 participants answered that they considered them when giving some of the answers and the remaining 10 responded that they did not consider machine estimates that much. On average, the predictions of the former group improved more with the user model (Figure 2).

The feedback the participants gave to the keywords differed between the two systems in several aspects. Table 2 lists average feedback values and machine estimates for 10 keywords with the lowest nominal p-values (two-sample t-test without assuming equal variances between the two systems), showing the keywords that were the most different between the systems. The average correlation to machine estimates for users in IE system was 0.46 (0.33 for the baseline system) and the average variance of feedbacks on keywords was 0.043 (0.060 for the baseline system). These support the hypothesis that the participants considered the machine estimates in the IE system. In the IE system, the average correlation to machine estimates after the correction done by the user model declined to -0.017 which suggests that the user model was successful in reducing the dependency to the training data.

DISCUSSION AND CONCLUSION

We described a user modelling methodology in probabilistic models for disentangling latent user knowledge from observed user feedback that was given in response to machine revealing information from the training data. We used this to guard against double use of the training data and overfitting in interactive machine learning. The proof-of-concept user study in interactive knowledge elicitation of feature relevance information in sparse linear regression showed the potential of the approach for improving prediction performance.

Keyword	*Machine*	*IE*	*Baseline*	*P-value*
best	0.89	0.90	0.80	0.014
great	1.00	0.95	0.83	0.031
disappointed	0.99	0.96	0.85	0.038
heavy	0.20	0.36	0.52	0.038
buy this	0.70	0.77	0.63	0.053
steel	0.19	0.12	0.24	0.073
line	0.34	0.15	0.06	0.088
don't	0.47	0.54	0.38	0.090
recommend	0.16	0.74	0.84	0.102
good	0.26	0.71	0.81	0.115

Table 2. Difference between user feedback in the baseline and IE systems (without user model).

Our approach is based on a simple rationality assumption of the user. It is, however, unreasonable to assume that users would in general behave completely rationally. In particular, the amount information and the way it is presented to the user are important factors that should be considered in designing interactive machine learning systems and user models. In our experiment, the interaction was based on reporting probability values through sliders. Such probability elicitation can be prone to the anchoring effect [5, 22]. Arguably, similar psychological mechanisms will appear in many interactive machine learning applications, since the machine often needs to guide the user for efficient interaction. This paper demonstrates that as long as the user modelling is able to capture the main sources of defects in the interaction, it would be expected to improve the results.

We believe that, as the field of interactive machine learning matures, more consideration will be put into the possible defects in the human–machine interaction. In particular, better user modelling will allow the user to behave naturally in providing feedback to the machine, while the machine (that is, the design of the machine learning system and models) will account for the inevitable human factors and biases [2, 5, 9, 12, 15, 17, 22, 24] to optimally combine the complementary knowledge and skills of the user and the machine. Our work is a step towards this with regard to overfitting in the interaction.

ACKNOWLEDGMENTS

This work was financially supported by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN; Grants 295503, 294238, 292334, and 284642), Re:Know funded by TEKES. We acknowledge the computational resources provided by the Aalto Science-IT Project. We thank Marta Soare for collaboration and helpful comments in early stage of the project.

REFERENCES

1. Homayun Afrabandpey, Tomi Peltola, and Samuel Kaski. 2017. Interactive Prior Elicitation of Feature Similarities for Small Sample Size Prediction. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 265–269. DOI: http://dx.doi.org/10.1145/3079628.3079698

2. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
3. Pedram Daee, Tomi Peltola, Marta Soare, and Samuel Kaski. 2017. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning* 106, 9 (2017), 1599–1620. DOI: http://dx.doi.org/10.1007/s10994-017-5651-7
4. Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 6248 (2015), 636–638.
5. Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. 2005. Statistical Methods for Eliciting Probability Distributions. *J. Amer. Statist. Assoc.* 100, 470 (2005), 680–701. DOI: http://dx.doi.org/10.1198/016214505000000105
6. Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 6245 (2015), 273–278.
7. José Miguel Hernández-Lobato, Daniel Hernández-Lobato, and Alberto Suárez. 2015. Expectation propagation in linear regression models with spike-and-slab priors. *Machine Learning* 99, 3 (01 Jun 2015), 437–487. DOI: http://dx.doi.org/10.1007/s10994-014-5475-7
8. Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (01 Jun 2016), 119–131. DOI: http://dx.doi.org/10.1007/s40708-016-0042-6
9. Tzu-Kuo Huang, Lihong Li, Ara Vartanian, Saleema Amershi, and Xiaojin Zhu. 2016. Active Learning with Oracle Epiphany. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2820–2828. http://papers.nips.cc/paper/6155-active-learning-with-oracle-epiphany.pdf
10. Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive Optimization for Steering Machine Classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1343–1352. DOI: http://dx.doi.org/10.1145/1753326.1753529
11. Josua Krause, Adam Perer, and Enrico Bertini. 2014. INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1614–1623.
12. Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3075–3084. DOI: http://dx.doi.org/10.1145/2556288.2557238
13. Luana Micallef, Iiris Sundin, Pekka Marttinen, Muhammad Ammad-ud din, Tomi Peltola, Marta Soare, Giulio Jacucci, and Samuel Kaski. 2017. Interactive Elicitation of Knowledge on Feature Relevance Improves Predictions in Small Data Sets. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 547–552. DOI: http://dx.doi.org/10.1145/3025171.3025181
14. Thomas Mühlbacher and Harald Piringer. 2013. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 1962–1971.
15. Edward Newell and Derek Ruths. 2016. How One Microtask Affects Another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3155–3166. DOI: http://dx.doi.org/10.1145/2858036.2858490
16. Daniel Russo and James Zou. 2016. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Vol. PMLR 51. 1232–1240.
17. Dominik Sacha, Michael Sedlmair, Leishi Zhang, John A Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C North, and Daniel A Keim. 2017. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* 268 (2017), 164–175.
18. Advait Sarkar, Mateja Jamnik, Alan F Blackwell, and Martin Spott. 2015. Interactive visual machine learning in spreadsheets. In *Visual Languages and Human-Centric Computing (VL/HCC), 2015 IEEE Symposium on*. IEEE, 159–163. DOI: http://dx.doi.org/10.1109/VLHCC.2015.7357211
19. Burr Settles. 2010. *Active learning literature survey*. Computer sciences technical report 1648. University of Wisconsin, Madison.
20. Iiris Sundin, Tomi Peltola, Muntasir Mamun Majumder, Pedram Daee, Marta Soare, Homayun Afrabandpey, Caroline Heckman, Samuel Kaski, and Pekka Marttinen. 2017. Improving drug sensitivity predictions in precision medicine through active expert knowledge elicitation. *arXiv preprint arXiv:1705.03290* (2017).
21. Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. 2009. EnsembleMatrix: Interactive Visualization to Support Machine Learning with Multiple Classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1283–1292. DOI: http://dx.doi.org/10.1145/1518701.1518895

22. Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. DOI: http://dx.doi.org/10.1126/science.185.4157.1124

23. Stef Van Den Elzen and Jarke J van Wijk. 2011. BaobabView: Interactive construction and analysis of decision trees. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*. IEEE, 151–160.

24. Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*.

Publication V

Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, Benjamin Blankertz. Integrating Neurophysiological Relevance Feedback in Intent Modeling for Information Retrieval. *Journal of the Association for Information Science and Technology*, 2019.

Integrating Neurophysiologic Relevance Feedback in Intent Modeling for Information Retrieval

Giulio Jacucci
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: giulio.jacucci@helsinki.fi

Oswald Barral
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: oswald.barral@helsinki.fi

Pedram Daee
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, P.O.Box 15400, Aalto FI-00076, Finland. E-mail: pedram.daee@aalto.fi

Markus Wenzel
Neurotechnology Group, Technische Universität Berlin, Berlin 10587, Germany. E-mail: markus.wenzel@hhi.fraunhofer.de

Baris Serim
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: baris.serim@helsinki.fi

Tuukka Ruotsalo
Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, P.O. Box 68, (Pietari Kalmin katu 5), Helsinki FI-00014, Finland. E-mail: tuukka.ruotsalo@helsinki.fi

Patrik Pluchino
Human Inspired Technology Research Centre, University of Padova, Via Luzzatti 4, Padova 35121, Italy. E-mail: patrik.pluchino@unipd.it

Jonathan Freeman
Goldsmiths, University of London, New Cross, London SE14 6NW, UK. E-mail: j.freeman@gold.ac.uk

Luciano Gamberini
Human Inspired Technology Research Centre, University of Padova, Via Luzzatti 4, Padova 35121, Italy. E-mail: luciano.gamberini@unipd.it

Samuel Kaski
Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, P.O.Box 15400, Aalto FI-00076, Finland. E-mail: samuel.kaski@aalto.fi

Benjamin Blankertz
Neurotechnology Group, Technische Universität Berlin, Berlin 10587, Germany. E-mail: benjamin.blankertz@tu-berlin.de

Received November 20, 2017; revised April 20, 2018; accepted October 17, 2018

 Published online Month 00, 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24161

The use of implicit relevance feedback from neurophysiology could deliver effortless information retrieval. However, both computing neurophysiologic responses and retrieving documents are characterized by uncertainty because of noisy signals and incomplete or inconsistent representations of the data. We present the first-of-its-kind, fully integrated information retrieval system that makes use of online implicit relevance feedback generated from brain activity as measured through electroencephalography (EEG), and eye movements. The findings of the evaluation experiment (N = 16) show that we are able to compute online neurophysiology-based relevance feedback with performance significantly better than chance in complex data domains and realistic search tasks. We contribute by demonstrating how to integrate in interactive intent modeling this inherently noisy implicit relevance feedback combined with scarce explicit feedback. Although experimental measures of task performance did not allow us to demonstrate how the classification outcomes translated into search task performance, the experiment proved that our approach is able to generate relevance feedback from brain signals and eye movements in a realistic scenario, thus providing promising implications for future work in neuroadaptive information retrieval (IR).

Introduction

Information retrieval systems are confronted with a difficult task; deriving a user's information needs from limited explicit user signals and use these to retrieve information matching those needs. Although modeling the data to be retrieved has witnessed dramatic advances during the last decades, understanding users' information needs is still based on rather simple user signals, such as queries, clicks, speech commands, or other explicit interactions. As a result, understanding information needs implicitly without disrupting the user has become a central research challenge in information retrieval (IR). Neurophysiologic measures are promising candidates for implicitly gathering relevance feedback, as they reflect the inner state of the user and can be collected unobtrusively at high throughput (Cowley et al., 2016; Eugster et al., 2016; Jacucci, Fairclough, & Solovey, 2015; Wenzel, Bogojeski, & Blankertz, 2017). Neurophysiologic signals hold a great potential for information retrieval as they provide a novel user signal revealing interests and relevance towards a diverse digital content as they happen when users are consuming digital information. Neurophysiologic signals also carry extraordinary practical promise as numerous types of wearable devices are rapidly becoming integral part of people's everyday life.

However, successful application of neurophysiologic measures in IR encounters a dual uncertainty problem: (a) noisiness and unknown causes of responses in neurophysiologic signals make it difficult to interpret them, a problem exacerbated by the lack of stimulus control in realistic settings, and (b) the IR process involves inherent uncertainty originating from the ambiguity and inconsistency of the representations of data to be retrieved. Unlike explicit relevance feedback that has low uncertainty due a user's overt control, implicit relevance feedback techniques are intrinsically noisy. When observing a user's click-through activity or brain responses to infer relevance feedback, the uncertainty of the feedback accuracies becomes higher, and incorporating this feedback within an interactive IR system requires novel computational solutions. The integration of brain signals has been especially challenging; even though they have shown promise, their use beyond laboratory experiments with very controlled stimuli remains largely unexplored. Previous work displays a limited number of unambiguous stimuli on the screen and/or constrains user interaction to decrease the amount of noise (Eugster et al., 2014; Eugster et al., 2016). In contrast, realistic search interfaces are characterized by dense information, potential ambiguity regarding the relevance of search results, and user interaction.

Our work provides the following contributions:

1. We demonstrate an approach able to predict implicit relevance feedback from human-brain measurements in a realistic search scenario.
2. We present a first-of-its-kind interactive IR system that combines brain-based feedback and eye tracking with scarce explicit feedback for improved relevance predictions.

The article is structured as follows. First, a brief discussion on related work on implicit relevance feedback in IR using brain–computer interfaces (BCIs) is presented. The section *An Approach for Single-Trial Relevance Computation in IR* investigates the challenge of decoding single-trial event-related potentials (ERP) that involve semantic interpretation of complex stimuli with large variability. We follow with a detailed proposal of a neurophysiologic approach for relevance computation, providing validation proof for the method, while highlighting potential challenges to be addressed when integrating relevance computation from brain signals in an IR system.

In the subsequent section, *Addressing Uncertainty in an Online Neuroadaptive System through Interactive Intent Modeling* we propose interactive intent modeling as a particular retrieval and ranking approach that facilitates the elicitation of explicit and implicit relevance feedback. Our approach in this respect is characterized by combining modeling of neurophysiologic response with modeling interactively intent in IR. In the section *An Experiment in Neuroadaptive Literature Search* we report the evaluation of our approach through findings from an experiment (N = 16) showing that we can predict neurophysiology-based relevance feedback in complex data domains and realistic search tasks and combine it with explicit relevance feedback in interactive intent modeling.

Related Work

Traditional relevance feedback techniques involve asking a user to provide explicit judgments on the information content. These has proven to be problematic because, in practice, users are reluctant to interrupt their search task to

DOI: 10.1002/asi

provide relevance feedback, even although they are aware that doing so would improve their search performance (Kelly & Fu, 2006). An important bottleneck of information seeking systems is that a considerable amount of user relevance feedback on retrieved items is needed to properly explore the large information space (Daee, Pyykkö, Glowacka, & Kaski, 2016). To overcome this challenge, previous approaches investigated "implicit relevance feedback" as indexed from search behavior from mouse and keyboard interaction data to understand a user's interests and personalize and rank search results (Kelly & Teevan, 2003). Other sources of implicit feedback include eye tracking to infer a user's interest through various metrics such as fixation count, dwell time, pupil size, and scan paths (e.g., Gwizdka, 2014; Oliveira, Aula, & Russell, 2009; Puolamäki, Salojärvi, Savia, Simola, & Kaski, 2005), analysis of user's facial expressions (e.g., Arapakis, Athanasakos, & Jose, 2010), physiologic responses (e.g., Barral et al., 2015, 2016), or a combination of these (e.g., Arapakis, Konstas, & Jose, 2009; Moshfeghi & Jose, 2013). Lately, brain signals have been identified as promising sources for implicit relevance feedback and information personalization (e.g., Eugster et al., 2014; Eugster et al., 2016; Golenia, Wenzel, & Blankertz, 2015).

IR is one of the fields that could profit from this direct access to the mental processes of the brain (Golenia et al., 2015; Gwizdka & Mostafa, 2015, 2017). First of all, mental processes can reveal information about relevance in response to particular information items thereby providing an effective way to elicit implicitly relevance feedback with great efficiency gains in being able to expose more items and collect relevance feedback without disrupting the user's search process. Second, mental processes and psychophysiologic states can be used to automatically annotate information such as news with affective or relevance response for future use and collaborative filtering (Barral et al., 2016; Barral, Kosunen, & Jacucci, 2017), and finally affective states as detectable from the brain can provide important context information for when or how to present information to user considering awareness, cognitive workload and other mental states. Research at the intersection between brain-computer interfaces (BCIs) and IR is still in an early stage, and appropriate neurophysiologic methods have to be matched with the appropriate paradigms for HCI in IR. Kauppi et al. (2015) studied magnetoencephalographic signals alone and in conjunction with gaze signals to provide relevance feedback in an image retrieval task by using a static image database. Similarly, Eugster et al. (2014) decoded the EEG with the objective of providing relevance feedback in a text retrieval task by using a static text data set. Other studies (Golenia et al., 2015; Golenia, Wenzel, Bogojeski, & Blankertz, 2018) demonstrated how the brain response to relevant versus irrelevant information can be harnessed to improve image searches in ambiguous search tasks. Recently research in the neurophysiologic correlates of relevance have been studied by Moshfeghi, Pinto, Pollick, & Jose, (Moshfeghi, Pinto, Pollick, & Jose, 2013) using functional Magnetic Resonance Imaging (fMRI) revealing three brain regions in the frontal, parietal and temporal cortex where brain activity differed between processing relevant and non-relevant documents. Not only where in the brain but also when relevance assessment phenomena happen have been studied for example by Allegretti et al. (2015) using a 64-channel EEG device. They found a significant variation between relevance and nonrelevance for the first 800 milliseconds (ms) of a relevance assessment process from the presentation of the image within the EEG signals. These studies are important as provide important additional evidence on the feasibility to include brain signal based relevance elicitation, however most studies focus on relevance of images and videos and less on text for which ca be more challenging to elicit and detect physiologic responses. Moreover, Eugster et al. (2016) gave relevant feedback on words from the Wikipedia database according to information extracted from EEG signals. The loop between brain and computer was closed by presenting new recommendations to the users according to the EEG-based feedback, which resulted in a significant information gain for about 70% of the participants of the study. This work constitutes presumably the first proof-of-concept IR systems that have performed automatic information filtering on the basis of brain activity alone.

Despite these advancement such as studies of neurophysiologic correlates of relevance, and applications using different kinds of stimuli, there is a lack of understanding on how to integrate neurophysiology-based relevance feedback in a realistic IR scenario. On one hand this includes the need of standardized approaches and procedures in research (Mostafa & Gwizdka, 2016) considering for example the use of machine learning. More importantly questions arise on what user intent and retrieval models are best suited to process the obtained implicit relevance feedback and how this can be combined with other relevance information obtained for example through explicit feedback.

An Approach for Single-Trial Relevance Computation in IR

Uncertainty in Single-Trial EEG Decoding

Because of the comparably high conductivity of the brain and scalp with respect to the one of the skull, electrical signals arrive spatially smeared at the EEG sensors, leading to low signal-to-noise ratio. Each sensor receives a mixture of signals from many sources in the brain and, conversely, the signals of one particular brain source are recorded at many different electrodes with a broad spatial profile. The predominant approach for real-time decoding is to employ multivariate data analysis methods from the field of machine learning (Lemm, Blankertz, Dickhaus, & Müller, 2011) and to train subject-specific decoding models on calibration data. Although this approach is comparably effective, a high degree of uncertainty in single-trial analysis remains, probably because of the very high number of potentially disturbing sources.

The perception and cognitive evaluation of visual stimuli, such as information presented on a computer screen, is reflected by event-related potentials (ERPs). In the well-known ERP-based *Row-Column Speller* (Farwell & Donchin, 1988), users concentrate on a target symbol while the rows and columns of the matrix of all symbols are flashing randomly. If the user fixates on the target symbol by gaze, the detection tasks boil down to a mere detection of flashes. More recent ERP-based spellers, such as the *Center Speller* (Treder, Schmidt, & Blankertz, 2011) circumvent the gaze-dependency of the *Row-Column Speller* by posing a higher load on the user as it requires the recognition of a target shape or color. Advancing further into the realm of IR (3), the evaluation of information involves semantic interpretation and more complex stimuli with large variability. In this escalation, the brain responses follow an increasingly less common temporal structure across trials. This leads to a larger variability in the latencies, but also in the morphology of the ERPs and, therefore, to a larger uncertainty in the decoding, see Figure 1.

The challenge of extracting information from a single-trial EEG gets even larger when free-viewing applications are considered. A suitable method for the investigation of free-viewing tasks are eye-fixation-related potentials (EFRP), see (Baccino & Manunta, 2005). Nevertheless, the decoding of the cognitive processes is hampered. On one hand, further unrelated brain activity connected to saccades and artifacts from eye movements overlay the EEG and, on the other hand, the temporal relationship between target-related ERP components and eye movements is variable because task-relevant processing of visual objects may already start before the beginning of a saccade, for example when the visual object is still at a peripheral location (Wenzel, Golenia, & Blankertz, 2016).

Neurophysiology-Based Relevance Computation

We propose a method to predict the relevance of textual keywords from brain signals and eye movements. The approach follows a supervised learning scheme, in which a user-specific classifier is trained by using labeled data. Then, the trained classifier can be used to generate relevance measures online, which can potentially be used in a feedback loop while the user interacts with the system. This machine learning approach is parallel to most modern BCI systems (Nijholt et al., 2008).

Training the classifier. The purpose of this first phase (referred as "the calibration phase") is to gather enough brain activity associated with the user's relevance judgments to train a classifier that will then be used to generate relevance measures online. A series of keywords for which relevance labels are known are presented to the user, and eye tracking is employed to identify when an eye fixation falls on a keyword. For each fixation that falls on a keyword, a high-dimensional feature vector is extracted from the EEG and eye movements (see below) and is labeled as "relevant" or "irrelevant" according to the known label of the keyword. A classification function is then trained to discriminate the feature vectors of the "relevant" and the "irrelevant" classes. To this end, regularized linear discriminant analysis is used (Friedman, 1989), whereby the shrinkage parameter is calculated with an analytic method (Ledoit & Wolf, 2004; Schäfer & Strimmer, 2005).

Online relevance computation. Once the system has been calibrated for the specific user by training a user-specific classifier, the user can interact with the system while EEG signals and eye movements are monitored (referred to as "the online phase"). For each keyword fixated on, a high-dimensional feature vector is extracted (see below), and the classifier infers its label online as belonging to the "relevant" or "irrelevant" classes. This means that the relevance predictions are available to the system in real time and can be used in an adaptive feedback loop.

Feature extraction. High-dimensional feature vectors are extracted from EEG channels recorded at 1000Hz according to the following steps: First, the multichannel EEG signal is re-referenced to the linked mastoids and low-pass filtered (with a second order Chebyshev filter; 42 Hz pass-band, 49 Hz stop-band). The continuous signal is then segmented by extracting the interval from 100 ms to 800 ms after the onset of every eye fixation. Slow fluctuations in the signal are removed by baseline correction (i.e., by

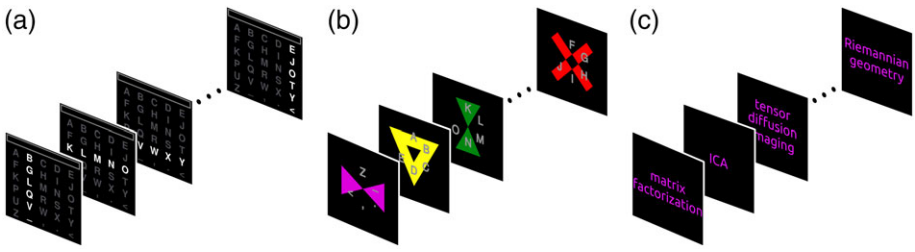

FIG. 1. From target to relevance detection. The classical row-column speller **(a)** which consists essentially in the detection of flashing. The center speller **(b)** relies on the recognition of a target shape/color. In contrast, the task to search for relevant terms **(c)** is incomparably more complex. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

subtracting the mean of the signal within the first 50ms after the fixation onset from each epoch). The signal is downsampled from the original 1000 Hz to 20 Hz to decrease the dimensionality of the feature vectors to be obtained (14 values per channel). A low dimensionality in comparison to the number of available samples has been shown to reduce the risk of overfitting to the training data, which in turn is beneficial for the classification performance (Blankertz, Lemm, Treder, Haufe, & Müller, 2011). The multichannel signal is vectorized by concatenating the values measured at the EEG channels at the 14 time points. The fixation duration is concatenated as an additional feature to the EEG feature vector. Other eye-tracking-related features (e.g., gaze velocity) are not considered as they are not provided in real time by the application programming interface of the device. Further, eye-movement-related signal components are not removed from the EEG because the classifier is expected to deal with task-unrelated eye-movements.

Method validation. To validate the approach in terms of computing relevance measures from semantic words, we carried out a *prior experiment* ($N = 15$). The main question addressed was whether relevance inference from the electroencephalogram (EEG) can be applied in settings where the interpretation of the semantics goes beyond the simple recognition of a previously known letter, picture, or shape that is repeatedly flashed. In the experiment, participants looked for words that belonged to semantic categories, and it was predicted in real-time which words, and thus which semantic category, was the one the user was interested in. Results showed that models using EEG features alone, and in combination with the eye fixation duration feature were able to generate single trial predictions on the keywords significantly above chance levels. Further, these predictions were aggregated in real time to provide reliable estimates of which were the semantic category of interest, showing slight improvements when adding fixation duration to the EEG-based feature vectors. Complete details on the *prior experiment* have been published separately in Wenzel et al. (2017).

The *prior experiment* provided several insights. First, it validated the use of EEG and eye gaze signals to infer subjective relevance of words that required interpretation with respect to their semantics in a free search task (as opposed to commonly used "counting" tasks). Further, predictions were generated on words that were presented simultaneously, relating neural activity to keywords using eye tracking. The *prior experiment* also evidenced the relatively low single-trial classification performances, which were successfully dealt with in real time by averaging over semantic categories. However, when interacting with a real IR system, the user interest and intentions may be more complex than as simulated in the *prior experiment*, and other mechanisms should be envisaged to integrate contextual information that may help to correct the noisy single-trial prediction accuracies.

Addressing Uncertainty in an Online Neuroadaptive System through Interactive Intent Modeling

A promising solution to cope with the uncertainty in the user's intent is interactive intent modeling (Ruotsalo, Jacucci, Myllymäki, & Kaski, 2015), where the potential search intentions of the user are represented and visualized as keywords, their relevance are estimated using feedback signals from the user, and information corresponding to the model is retrieved. In terms of neuroadaptive systems, intent modeling can mitigate both the uncertainty related to the noise present in neurophysiologic signals and the mismatch between the user's articulation of information needs and the encodings of the information to be retrieved.

Adapting the intent model from suboptimal and noisy user feedback

The intent model directly couples the potentially suboptimal user feedback originating from implicit and explicit user signals. The implicit feedback is connected to explicit feedback by considering source-specific probabilistic assumptions on their uncertainties. This provides the flexibility to learn the true uncertainty of each feedback given all preceding feedback.

Estimating the intent model. The relevance of keywords in the model is described with a linear Gaussian model, with which the accuracy of the feedback may differ for the different source types (implicit or explicit). The relevance of keyword i is modeled as

$$y_i \sim N\left(x_i\phi, \sigma^2/w_i\right), \tag{1}$$

where x_i is the feature vector representing that keyword, ϕ is the unknown weight vector which is shared between all keywords and maps the feature vectors to relevance values representing user intent, σ^2 is the variance of feedback noise, and w_i models the accuracy of the relevance feedback. We assume prior distributions on the parameters to be

$$\phi \sim N(0, \lambda I),$$
$$\sigma^2 \sim InverseGamma(\alpha_{\sigma^2}, \beta_{\sigma^2}),$$
$$w_i \sim Gamma(\alpha_w, \beta_w),$$

where λ, α_{σ^2}, and β_{σ^2} are fixed hyperparameters. A key aspect of our approach is that we distinguish between implicit and explicit feedback by using different hyperparameters for prior of the accuracy values, that is, $\left(\alpha_w^{exp}, \beta_w^{exp}\right)$ for explicit feedback and $\left(\alpha_w^{imp}, \beta_w^{imp}\right)$ for implicit feedback.

The posterior of the model estimates both the user's current search intent (ϕ) and the accuracy of the user relevance feedback (w_is). As mentioned, the accuracies of the user feedback on keywords are unknown and drawn from

DOI: 10.1002/asi

a gamma distribution with two parameters: alpha and beta. The model differentiates among explicit and implicit feedback by using different sets of hyper-parameters for the gamma distribution. The explicit feedback is considered very certain (a gamma distribution with mean 1 and very small variance, that is, $\alpha_w^{exp} = 100, \beta_w^{exp} = 100$). On the other hand, the implicit feedback is uncertain *a priori* (gamma distribution with mean 0.5 and large variance, that is, $\alpha_w^{imp} = 1, \beta_w^{imp} = 2$), and therefore, its accuracy is mostly inferred from observations. For example, if the implicit feedback is in line with the previous history of feedback, then it will be inferred as certain and will contribute to the user model. However, if it contradicts the system's current belief, learned from sequence of feedback, then its accuracy may be inferred as a low value and it will not affect the user model (the posterior of ϕ) much. The model infers the true accuracies and corrects the noise in the feedback. We use mean-field variational inference for the posterior inference (Attias, 1999; Kangasrääsiö, Chen, Glowacka, & Kaski, 2016).

Estimating Document Relevance

In addition to estimating the relevances for the keywords in the intent model, the relevances of the documents are estimated and ranked. We employ the feature transformation that projects the relevances estimated for the keywords to the documents (Daee et al., 2016). The underlying principle is that the transformation projects documents in the feature space of the keywords as the relevance of a document is a weighted sum of the relevance of individual keywords that have appeared in it. Based on this projection, the relevance of a document also follows Equation 1 with the difference that the document feature vector is generated from the feature projection.

Exploring uncertainty. Estimating the intent model by directly exploiting the feedback observed from the user yields to showing items like those already judged relevant by the user in the previous iterations. Because the implicit feedback observed from the user may be inaccurate, this exploitative choice might cause the intent model to converge to a suboptimal representation of the user's intention. Alternatively, the system might exploratively select items that are relevant, but also uncertain. These items are likely to be better for obtaining feedback in subsequent iterations as they are novel and not too similar to the ones already judged by the user.

Multiarmed bandits have been shown to be able to model this exploration and exploitation dilemma in information seeking (Ruotsalo et al., 2015). We use the Thompson sampling algorithm (Agrawal & Goyal, 2013) as a solution to the multiarmed bandit problem, to control the exploration and exploitation balance of the recommended keywords and documents (Daee et al., 2016). The idea behind Thompson sampling is that the uncertainty in the marginal posterior of ϕ can by itself control the exploration and exploitation of the items. To implement the algorithm, it is enough to draw a sample from the posterior and rank all the keywords and documents accordingly. In detail, the Thompson sampling algorithm performs the following steps in each iteration:

1. Draw a sample from the marginal posterior of ϕ and denote it as ϕ^p.
2. Rank all the keywords based on the inner product $x_i^T \phi^p$.
3. Rank all the documents based on the inner product $x_j^T \phi^p$.
4. Recommend the highest ranked items and gather the feedback.
5. Update the posterior.

Here, x_i and x_j denote the feature vectors of keyword i and document j (after the transformation) respectively. The highest ranked recommendations were expected to consider the balance between exploration and exploitation (Agrawal & Goyal, 2013).

Visualizing the Intent Model for Explicit and Implicit Interaction

To enable implicit and explicit feedback from the user, the intent model needs to be visualized for interaction. The implicit feedback is captured via capturing eye fixations and EEG signal.

Interface views. The interface consists of two separate views: intent model view and document view. The intent model view, shown in Figure 2, visualizes the top-k keywords chosen based on their estimated weights resulting from the Thompson sampling algorithm. The view employs a circular layout chosen to increase eye tracking accuracy, which is higher at the center of the screen. The keyword are positioned randomly but the layout is optimized to increase the distance between neighboring keywords for more robust matching with eye fixations. The document view, shown in Figure 3, has a conventional ranked list visualization.

Interaction. The search is initiated by entering a query, which results in the first set of results retrieved by the system. To direct the search, users can open a view that displays a set of keywords that are potentially relevant to the users' search intent. The users can examine these keywords and provide explicit relevance feedback on one of the keywords by clicking on it. Although users examine the keywords, the physiologic classifier generates implicit relevance feedback on them. The system then updates the intent model by taking into account both the explicit relevance feedback, and the implicit feedback generated from the keywords the user fixated on. The system then returns the next iteration of results. This process is repeated until the user decides to change the query or ends the search task. Figure 4 depicts the user-system interaction as a control loop.

DOI: 10.1002/asi

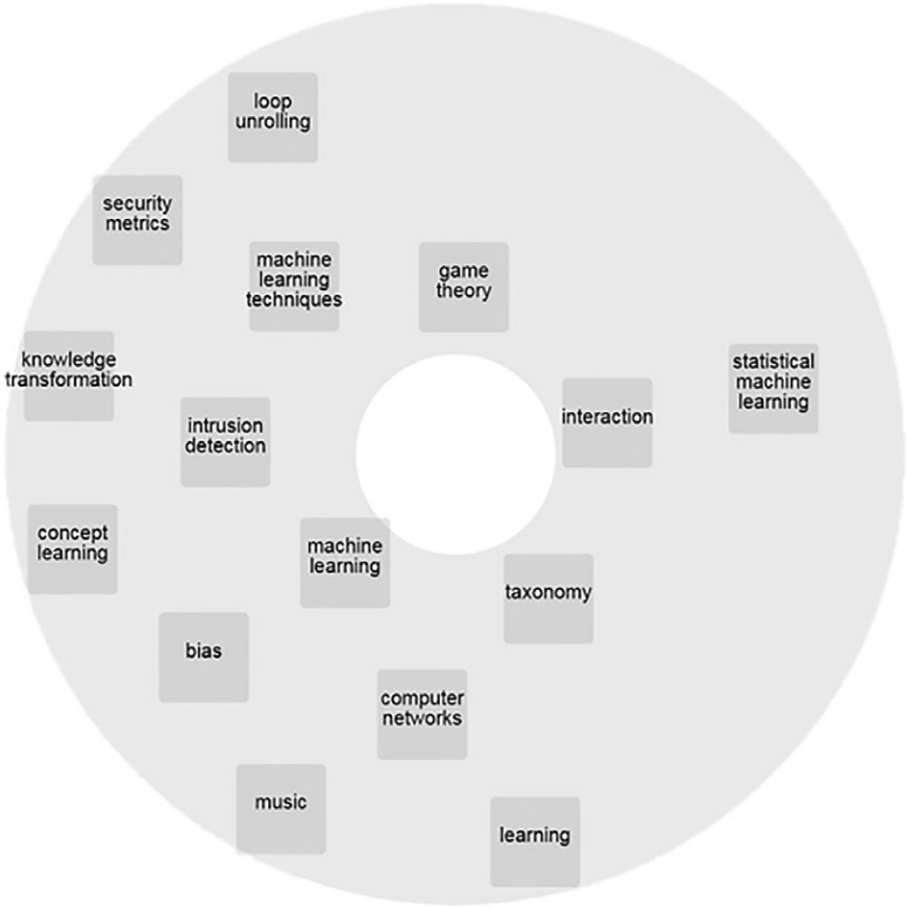

FIG. 2. A screenshot of the user interface displaying the intent model view.

An Experiment in Neuroadaptive Literature Search

This experiment helps to evaluate the approach and system presented in the previous two sections by investigating the following questions:

Is it possible to predict online relevance from neurophysiology in a realistic search task and integrate it as implicit feedback in combination with explicit feedback in interactive intent modeling?

System Apparatus

The system that integrates neurophysiology-based implicit feedback with interactive intent modeling is implemented as a web application using a frontend (the *interface*) - backend (the *engine*) architecture, see Figure 5. The engine comprises of three main components: The *Controller*, which coordinates the different components of the system; the *Physiologic Classifier*, which generates real-time implicit relevance feedback, and the *Interactive Intent Model*, which handles the user model and the information items of the system. The *Physiologic Classifier* is implemented within the framework of the BBCI-Toolbox.[1] For each gaze-fixation, the classifier sends to the *Controller* a relevance value. The *Controller* checks whether the fixation falls on a keyword visible on the screen to associate the predicted relevance value to it. For collecting eye movements, the system uses the SensoMotoric Instruments RED500 eye tracker, interfaced through the SMI iViewX SDK.[2] For collecting brain signals, the system supports the BrainProducts QuickAmp and BrainAmp amplifiers,[3] both of which recorded 32 EEG channels at a sampling rate of 1000 Hz. The *Interactive Intent Model* uses the same document-retrieval model as in Ruotsalo et al. (2013) to select subset of documents, and uses a data set from the following data sources: the Web of Science prepared by Thomson Reuters, Inc., the digital library of the Institute of Electrical and Electronics Engineers (IEEE), the digital library of the Association of Computing Machinery (ACM), and the digital library of Springer. The hyperparameters of the intent model were tuned as $\alpha_{\sigma^2} = 2$, $\beta_{\sigma^2} = 0.1$, and $\lambda = 0.1$ based on pilot experiments (N = 27).

Participants

Sixteen participants (3 females) took part in the experiment. The participants ranged from 22 to 39 years old

[1] https:/github.com/bbci/bbci_public

[2] http://www.smivision.com/

[3] http://www.brainproducts.com/

DOI: 10.1002/asi

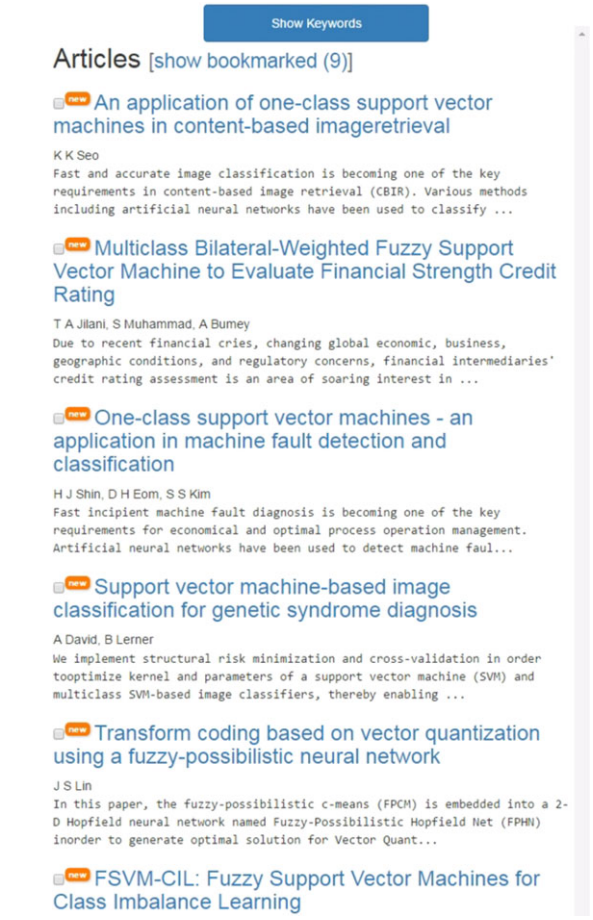

FIG. 3. A screenshot of the user interface displaying the document view. [Color figure can be viewed at wileyonlinelibrary.com]

(M = 28.3). Three participants were postdoctoral researchers, and the rest were students (8 post-graduate, 5 undergraduate) from the University of Helsinki in Finland and the University of Padova in Italy. The participants reported themselves as being physically and mentally healthy. The participants reported a good level of English (M = 4.0, SD = 0.9, on a 1 to 5 scale) and high expertise in computer science (M = 4.4, SD = 0.6, on a 1 to 5 scale). Their experience with browsing scientific literature (M = 3.6, SD = 0.9, on a 1 to 5 scale) and their prior knowledge of machine learning (M = 2.8, SD = 1.5, on a 1 to 5 scale) varied.

Procedure and Experimental Task

At the beginning of the session, the participants were welcomed and briefed as to the procedure and purpose of the experiment before signing the informed consent form. The participants were instructed about the duration of the experiment and reminded that they could withdraw from the experiment at any point in time, without facing negative consequences. Although the physiologic sensors were set up, the participants filled a background information questionnaire. Following, a standard 9-point eye tracker calibration procedure was carried out repeatedly until reaching an error smaller than 0.5 degrees of visual angle.

The calibration phase. The participants then engaged in the calibration phase for around 1 hour, until the system had collected enough data points to train the physiologic classifier. The participants were allowed to have small breaks during the calibration phase whenever they felt tired or their concentration was diminishing. To collect training data for the physiologic classifier, we generated a data set that matched the application domain by using a subset of the data set used by the interactive intent model system. The data set consisted of a set of topics with associated keywords and was created using expert judgments in an iterative process that aimed at minimizing the overlaps between the topics, while maximizing the dissociation between relevant and irrelevant keywords to a given topic.[4]

Participants were prompted with a list of five topics, randomly selected from the calibration data set. On selecting a topic, a series of keywords were shown to the user, who was asked to select the keywords relevant to the topic. This procedure was repeated iteratively for several topics, until the system had gathered enough data to train the physiologic classifier.[5]

The online phase. Once enough data had been collected and the physiologic classifier had been trained, the participants engaged in the online phase. Participants were provided the following instructions:

Imagine that you are going to write an essay about topic X. Please bookmark the articles on the scroll list that you think are relevant to the topic, so that you can use them later in the essay. You will later be asked to write a short outline of the essay based on your bookmarked articles.

The participants had to perform two versions of the same task, using the topics "neural networks" and "support vector machines." One of the tasks was performed using the full system. The other task was performed using a baseline system, which behaved in the exact same way as the full system, but no implicit relevance feedback was fed to the interactive intent model system. Instead, only the explicit feedback provided by the user was used to refine the user model and present the next iteration of results. The participants were unaware that they were using two different systems, and they were naïve about the systems' implementation.

For evaluation purposes, the participants were prompted at the end of each iteration with a dialog asking them to label the relevance of the keywords they had fixated on (on a scale from 0 to 5). This allowed the "ground truth" to be collected on the relevance of the presented keywords as perceived by the users. This was otherwise not

[4] For review: Refer to *Appendix A* for more details on the generation of the calibration data set.

[5] For review: Refer to *Appendix B* for details on how the assessment of keywords' relevance was carried out by the participants during the calibration phase.

DOI: 10.1002/asi

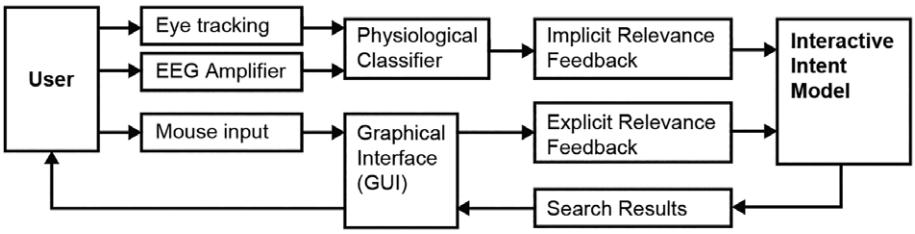

FIG. 4. Summary of the system as a control loop during the online phase.

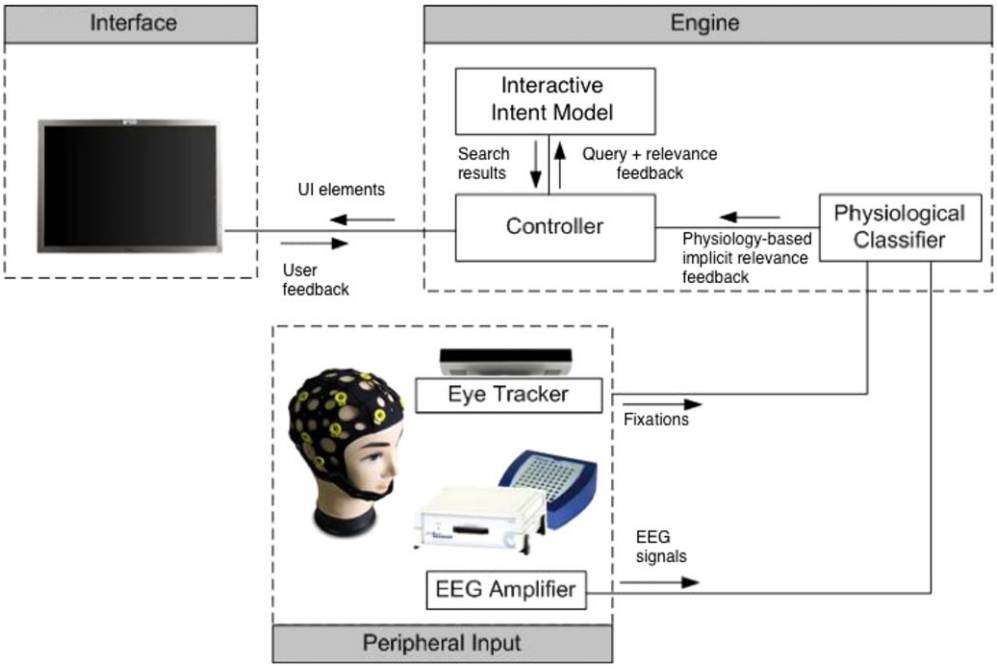

FIG. 5. Components of the system. [Color figure can be viewed at wileyonlinelibrary.com]

available, as the keywords were generated in real-time from the interactive intent model system, and their relevance naturally depends on the users' information needs, which were not known *a priori.*

The participants performed each task in the online phase for around 20 minutes, for a maximum of 10 iterations. The task and system type were counterbalanced. On completion of the task, participants were rewarded with two movie tickets. In total, the experiment lasted approximately 2.5 hours.

Measures and Analyses

Calibration phase. To evaluate the feasibility and performance of the system in predicting relevance from brain signals, we first evaluated the classification performance in the calibration phase. The data used in the calibration phase were controlled and had the advantage that the same data set was used to train the different user-specific classification models. Classification performance was computed in terms of area under the ROC curve (AUROC) and was evaluated using a standard 10 × 10 fold cross validation approach. AUROC is a widely used and sensible measure, even under class imbalances, that links the *true positive rate* and the *false positive rate* while avoiding possible misinterpretations such as the accuracy paradox (Zhu & Davidson, 2007).

To quantify the significance and the effect sizes of the implicit relevance feedback from the brain signals, we compared the classification performances against performances from prediction models learned from randomized labels. Standard permutation tests were applied for significance testing (Good, 2013). In detail, for each of the 16 participants, we ran within-participant permutation tests with 1000 iterations. For each iteration, we learned a classification model using randomized labels, and we then computed the p-value as the percentage of random classification performances that were equal to or greater than the true classification performance.

Online phase. The aim was to assess how well the classification performance achieved in the calibration phase transferred to the online phase, during which the users were engaged in a realistic information-seeking task, and

DOI: 10.1002/asi

the data presented to the user from which implicit relevance feedback was classified were generated in real-time. To do so, for each participant we computed the classification performance in terms of AUROC for each of the fixated keywords in the online phase in the tasks for which the participants used the full system. We used the feedback provided by the participants on the keywords as the labels. We binarized the user feedback, so that keywords that were rated between 0 and 2 were considered irrelevant and keywords that were rated between 3 and 5 were considered relevant. Further, participants *P05* and *P06* had to be rejected from the analysis because the server hosting the interactive intent model system went down during the execution of the online phase.

As explained in Section *Addressing Uncertainty in an Online Neuroadaptive System through Interactive Intent Modeling*, in each iteration, the intent model learns the relevance of all keywords from the available sequence of explicit and implicit feedback. Accordingly, we also computed the classification performance in terms of the AUROC of the relevance of keywords estimated by the intent model. This is the performance after the user model has accounted for the noise in implicit relevance feedback values coming from the physiologic classifier.

Task performance. After completion of the search task, participants were asked to write down some of the concepts that they had learned about the topics, which lead to a very heterogeneous collection of "mini-essays" not suited for comparison across participants. Instead, to assess whether using physiology-based implicit relevance measures had an influence on the task performance, we compared the quality of the documents that participants bookmarked when using the full system (including implicit relevance feedback) and when using the baseline system (that did not include implicit relevance feedback). In total, participants generated 397 bookmarks on 277 different documents. On the population level, documents had often been bookmarked using both system types (e.g, one participant had bookmarked a document in the baseline system condition, while another participant had bookmarked the same document in the full system condition). To assess any change in task performance between the two conditions, we therefore limited the analysis to a "representative" subset of documents that minimized overlaps. Documents were selected as "representative" for one of the conditions if on the population level, the document was bookmarked two or more times than in the other condition. This lead to a subset of 21 documents (8 representing the full system condition, and 13 representing the baseline system condition). Documents were then rated by 3 experts (on a 1-6 rating scale), on their *relevance* (i.e., is this document relevant to the search task), *obviousness* (i.e., is this a well-known overview article in a given research area), and *novelty* (i.e., is this article uncommon yet relevant to a given topic or specific subtopic in a given research area) (Ruotsalo et al., 2013). Ratings were averaged across experts, and Wilcoxon rank-sum tests were used to test for statistical differences between the two conditions (full system vs. baseline system), for each of the three rating categories (*relevance*, *obviousness*, and *novelty*).

Results

Calibration phase. Classification performance proved to be significantly better than random for 13 out of 16 participants, meaning that we were able to successfully train the classifier for around 80% of the participants. On the population level, AUROC resulted in 0.61 $\pm$ 0.02 (mean $\pm$ standard error of the mean). Figure 6 presents the individual classification performances in the calibration phase.

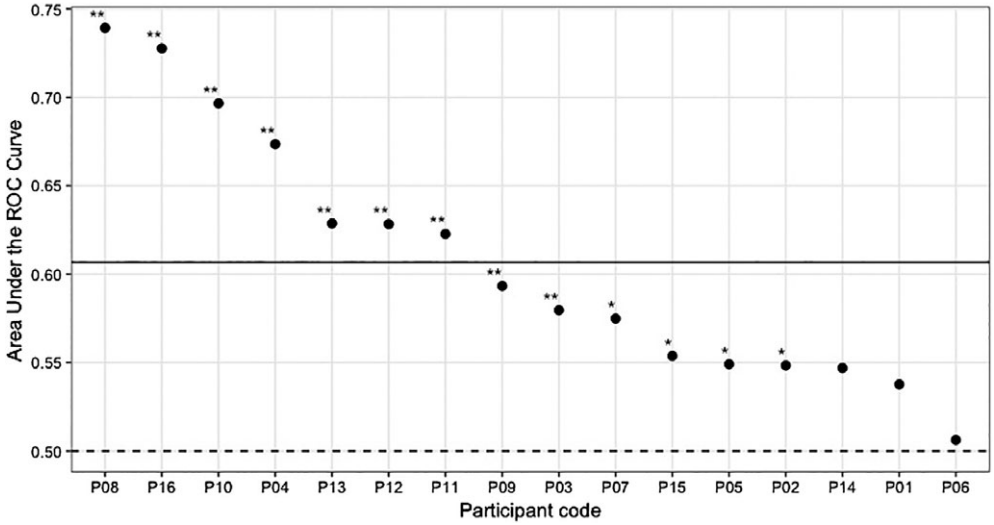

FIG. 6. Individual classification performances in the calibration phase in terms of area under the ROC curve (AUROC), and improvement over the random baseline at the levels of $p < 0.05$ (*), and $p < 0.001$ (**). The horizontal lines represent the mean (solid) and random (dashed).

DOI: 10.1002/asi

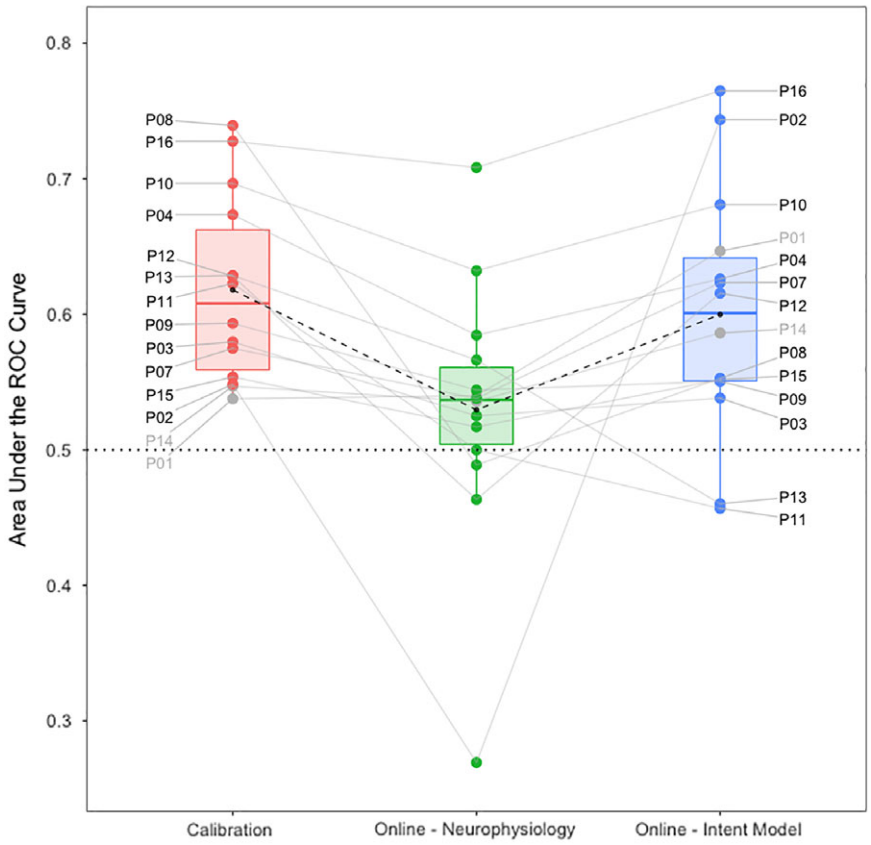

FIG. 7. Individual classification performance in terms of area under the ROC curve (AUROC). Left: offline prediction in the "calibration phase." Middle: neurophysiologic prediction in the "online phase." Right: intent model prediction in the "online phase." Smaller black dots and dashed lines indicate mean classification performance. The dashed horizontal line represents random classification. Participants for which calibration did not outperform random predictions are presented in gray. [Color figure can be viewed at wileyonlinelibrary.com]

Online phase. Online relevance predictions as directly obtained through the physiologic classifier presented averaged AUROC values on the population level of 0.53 ± 0.03 (mean ± standard error of the mean). The performance was improved by the user model, leading to averaged AUROC values of 0.60 ± 0.03. In fact, the intent model increased prediction performance for 10 out of the 12 participants for which a classifier was successfully trained in the calibration phase, representing over 80% of these participants. Figure 7 shows the results of the classification performance for the calibration phase and for the online phase, in terms of the implicit relevance feedback, both as directly obtained through classification of brain signals, and as inferred by the intent model.

Task performance. Wilcoxon rank-sum tests did not show statistical difference between the full system and baseline system, for any of the rating categories: In terms of *relevance*, expert ratings provided to representative documents of the full system (*Mdn* = 3.5) did not significantly differ from those of the baseline system (*Mdn* = 4.67), $W = 69$, $p = 0.22$. In terms of *obviousness*, expert ratings provided to representative documents of the full system (*Mdn* = 2.67) did not significantly differ from those of the baseline system (*Mdn* = 3.33), $W = 73.5$, $p = 0.12$. In terms of *novelty*, expert ratings provided to representative documents of the full system (*Mdn* = 3.83) did not significantly differ from those of the baseline system (*Mdn* = 3.67), $W = 55.5$, $p = 0.82$.

Discussion and Conclusions

A methodology for predicting implicit relevance feedback from human-brain measurements in a realistic search scenario was presented. The methodology was implemented in a first-of-its-kind interactive IR system that combines brain-based feedback and eye tracking with scarce explicit feedback. To our knowledge, the presented system is the first closed-loop IR system that utilizes brain-based feedback, combines it with eye-tracking and explicit feedback, and is evaluated in realistic IR tasks.

Empiric Findings

The empiric evidence suggests that the presented methodology allows to reliably train classification models for implicit relevance prediction by using complex real-world data. The results show that the classification performance significantly outperforms random predictions for over 80%

DOI: 10.1002/asi

of the participants, with some of the participants reaching AUROC values over 0.7. One explanation for the random classification outcomes among the remaining approximately 20% of participants could be the fact that BCI control does not work for a non-negligible proportion of users (approximately 15 - 30%) (Acqualagna, Botrel, Vidaurre, Kübler, & Blankertz, 2016; Allison et al., 2010; Blankertz et al., 2010; Guger et al., 2009). These results are comparable to the ones obtained in the *prior experiment* (see Section *Validating the Relevance Computation Method*, and (Wenzel et al., 2017)), where a limited and controlled data set of keywords was used.

In addition, the results show that the classification performances achieved using the controlled "calibration data set" in the calibration phase transferred to the online phase, during which the retrieved documents and keyword varied for each participant, and their perception of relevance was related to their current information needs, rather than to a predefined experimental task. Although the classification performance decreased as expected, the overall distribution across participants remained above random classification levels.

Further, we demonstrate that the approach is able to combine the noisy neurophysiology-based implicit relevance feedback with limited explicit feedback (one per search iteration), which improved the classification performance for over 80% of the participants for which we had successfully trained a classifier during the calibration phase.

Figure 7 shows atypical values for participant *P02*. By looking at the data, we found out that this participant provided highly unbalanced ground truth in the online phase (i.e., 96% of the ground truth provided was from the relevant class), which explains the drastic changes in the AUROC values. Thus, the magnitude of such changes in the performance measures should be interpreted cautiously. Further, we looked at the participants who consistently presented high AUROC values (i.e., *P04*, *P10*, and *P16*) to identify factors that could explain their better performance, which could potentially be used to improve the overall model and the design of future studies (e.g., level of education, prior knowledge about the topic, reported satisfaction and engagement with the system, etc.). We did not find anything especially noteworthy about them, nor performance differences between undergraduate and postgraduate participants.

Limitations

Our approach and study includes at least the following limitations. First, the predicted relevance from physiology, although promising, still leaves room for improvement, both in terms of classification performance and uniformity across participants. Second, as the online phase involves an online interaction loop between the participants and the system, we could not perform offline permutation tests to evaluate how much did the achieved classification performances improved over random classifications (as done for the calibration phase). This would have provided further empiric insight on how much the classification performance transferred from the controlled calibration phase to the realistic online phase, as well as on how much of the performance of the intent model is explained by the implicit feedback and how much by the scarce explicit feedback. Third, the analysis on the selection behavior of bookmarked documents did not yet yield conclusive results in terms of task performance improvements. Future work should extend the presented results by further studying how the reported classification performances could transfer over to search task performance. Overall, although we report on the first-of-its-kind closed-loop information retrieval system that fully integrates neurophysiologic signals while users perform real information seeking task, the experimental method and results indicate that there is still room for improvement, both to demonstrate the impact of the implicit feedback to the overall intent model, as well as to exemplify how using neurophysiologic input transfers to improved task performance.

Implications and Future Work

In essence, relevance judgments happen in the human brain and therefore the most intriguing way to predict relevance is to directly utilize the brain signals. These signals have advantages over the more conventional sources of user signals from a practical IR point of view. The recording of the relevance judgments directly from human neurophysiology do not require any explicit user interaction, such as user actively clicking on items. The current work contributes showing that predicting the relevance from neurophysiology on information presented in realistic information retrieval system responses is possible with promising accuracy. Moreover, it is demonstrated that the relevance prediction methodology can be operationalized as a part of a closed-loop information retrieval system. In concrete the work contributes not only a reference approach and procedure but proposes interactive intent modeling (Ruotsalo et al., 2015) as a promising user intent and retrieval model suited for processing implicit relevance and combining it with other relevance information such as explicit feedback. Our findings open a horizon for information retrieval systems that can detect relevance directly from human neurophysiology and combine this with potentially scarce explicit signals without requiring users to devote attention for laborious explicit interaction. Future studies can put more emphasis on demonstrating the actual task performance improvement that can be obtained, and can devise ways to collect repeated measures of implicit responses to reduce uncertainty either from the same participant or across participants. The inherent uncertainty in both the brain measurements, attentional focus of the user, and in data representations, however, call for computational methods for simultaneously modeling cognitional states and the data for which these states are associated with. Our findings already show a path towards closed-

loop systems that are able to analyze and utilize relevance and human cognition directly from wearable sensors as it is manifested as a part of human information search activities.

Acknowledgments

We thank Mats Sjöberg, Antti Kangasrääsiö, Nishadh Aluthge, and Hassan Abbas for their hard work in implementing the system and running experimental studies. This work has been supported by the European Commision (MindSee FP7-ICT; Grant Agreement #611570).

References

Acqualagna, L., Botrel, L., Vidaurre, C., Kübler, A., & Blankertz, B. (2016). Large-scale assessment of a fully automatic co-adaptive motor imagery-based brain computer interface. PLoS One, 11(2), e0148886. https://doi.org/10.1371/journal.pone.0148886

Agrawal, S., & Goyal, N. (2013). Thompson Sampling for Contextual Bandits with Linear Payoffs. In Proceedings of the 30th International Conference on Machine Learning (pp. 127–135). PMLR.

Allegretti, M., Moshfeghi, Y., Hadjigeorgieva, M., Pollick, F. E., Jose, J. M., & Pasi, G. (2015). When relevance judgement is happening?: An EEG-based study. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 719–722).

Allison, B., Luth, T., Valbuena, D., Teymourian, A., Volosyak, I., & Graser, A. (2010). BCI demographics: How many (and what kinds of) people can use an SSVEP BCI? IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18(2), 107–116.

Arapakis, I., Athanasakos, K., & Jose, J.M. (2010). A comparison of general vs personalised affective models for the prediction of topical relevance. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 371–378). New York, NY: ACM. https://doi.org/10.1145/1835449.1835512

Arapakis, I., Konstas, I., & Jose, J.M. (2009). Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In Proceedings of the 17th ACM International Conference on Multimedia (pp. 461–470). New York, NY: ACM. https://doi.org/10.1145/1631272.1631336

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In Proceedings of the fifteenth conference on uncertainty in artificial intelligence (pp. 21–30). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Baccino, T., & Manunta, Y. (2005). Eye-fixation-related potentials: Insight into parafoveal processing. Journal of Psychophysiology, 19(3), 204–215.

Barral, O., Eugster, M.J., Ruotsalo, T., Spapé, M.M., Kosunen, I., Ravaja, N., … Jacucci, G. (2015). Exploring peripheral physiology as a predictor of perceived relevance in information retrieval. In Proceedings of the 20th International Conference on Intelligent User Interfaces (pp. 389–399). New York, NY: ACM. https://doi.org/10.1145/2678025.2701389

Barral, O., Kosunen, I., & Jacucci, G. (2017). No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment. ACM Transactions on Computer-Human Interaction, 24(6), 40. Retrieved from http://doi.acm.org/10.1145/3157730. https://doi.org/10.1145/3157730

Barral, O., Kosunen, I., Ruotsalo, T., Spapé, M.M., Eugster, M.J.A., Ravaja, N., … Jacucci, G. (2016). Extracting relevance and affect information from physiological text annotation. User Modeling and User-Adapted Interaction, 26(5), 493–520. https://doi.org/10.1007/s11257-016-9184-8

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components – A tutorial. NeuroImage, 56(2), 814–825. https://doi.org/10.1016/j.neuroimage.2010.06.048

Blankertz, B., Sannelli, C., Halder, S., Hammer, E.M., Kübler, A., Müller, K.-R., … Dickhaus, T. (2010). Neurophysiological predictor of SMR-based BCI performance. NeuroImage, 51(4), 1303–1309. https://doi.org/10.1016/j.neuroimage.2010.03.022

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., … Jacucci, G. (2016). The psychophysiology primer: A guide to methods and a broad review with a focus on human–computer interaction. Foundations and Trends[textregistered] in Human—Computer Interaction, 9(3-4), 151–308. Retrieved from https://doi.org/10.1561/1100000065. https://doi.org/10.1561/1100000065

Daee, P., Pyykkö, J., Glowacka, D., & Kaski, S. (2016). Interactive intent modeling from multiple feedback domains. In Proceedings of the 21st International Conference on Intelligent User Interfaces (pp. 71–75). New York, NY: ACM. https://doi.org/10.1145/2856767.2856803

Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., Barral, O., Ravaja, N., Jacucci, G., & Kaski, S. (2016). Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. Scientific Reports, 6, 38580. https://doi.org/10.1038/srep38580

Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., Kosunen, I., Barral, O., Ravaja, N., … Kaski, S. (2014). Predicting term-relevance from brain signals. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 425–434). New York, NY: ACM. https://doi.org/10.1145/2600428.2609594

Farwell, L.A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and clinical Neurophysiology, 70(6), 510–523.

Friedman, J.H. (1989). Regularized discriminant analysis. Journal of the American Statistical Association, 84(405), 165–175. https://doi.org/10.1080/01621459.1989.10478752

Golenia, J.-E., Wenzel, M.A., & Blankertz, B. (2015). Live demonstrator of EEG and eye-tracking input for disambiguation of image search results. In Symbiotic interaction (pp. 81–86). Cham: Springer. https://doi.org/10.1007/978-3-319-24917-9_8

Golenia, J.E., Wenzel, M.A., Bogojeski, M., & Blankertz, B. (2018). Implicit relevance feedback from electroencephalography and eye tracking in image search. Journal of Neural Engineering, 15(2), 026002. https://doi.org/10.1088/1741-2552/aa9999

Good, P. (2013). Permutation tests: a practical guide to resampling methods for testing hypotheses. Berlin/Heidelberg, Germany: Springer Science & Business Media.

Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., … Edlinger, G. (2009). How many people are able to control a p300-based brain–computer interface (BCI)? Neuroscience Letters, 462(1), 94–98.

Gwizdka, J. (2014). Characterizing relevance with eye-tracking measures. In Proceedings of the 5th Information Interaction in Context Symposium (pp. 58–67). New York, NY: ACM. https://doi.org/10.1145/2637002.2637011

Gwizdka, J., & Mostafa, J. (2015). NeuroIR 2015: SIGIR 2015 workshop on neuro-physiological methods in IR research. In ACM SIGIR Forum (Vol. 49, pp. 83–88). ACM. https://doi.org/10.1145/2888422.2888435

Gwizdka, J., & Mostafa, J. (2017). NeuroIIR: Challenges in bringing neuroscience to research in human-information interaction. In Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17 (pp. 437–438). New York, NY: ACM Press. https://doi.org/10.1145/3020165.3022165

Jacucci, G., Fairclough, S., & Solovey, E.T. (2015). Physiological computing. Computer, 48(10), 12–16. Retrieved from http://ieeexplore.ieee.org/document/7310960/. https://doi.org/10.1109/MC.2015.291

Kangasrääsiö, A., Chen, Y., Glowacka, D., & Kaski, S. (2016). Interactive modeling of concept drift and errors in relevance feedback. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (pp. 185–193). New York, NY: ACM. https://doi.org/10.1145/2930238.2930243

Kauppi, J.-P., Kandemir, M., Saarinen, V.-M., Hirvenkari, L., Parkkonen, L., Klami, A., … Kaski, S. (2015). Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. NeuroImage, 112, 288–298. https://doi.org/10.1016/j.neuroimage.2014.12.079

Kelly, D., & Fu, X. (2006). Elicitation of term relevance feedback: An investigation of term source and context. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 453–460). New York, NY: ACM. https://doi.org/10.1145/1148170.1148249

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. SIGIR Forum, 37(2), 18–28. https://doi.org/10.1145/959258.959260

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. Neuroimage, 56(2), 387–399.

Moshfeghi, Y., & Jose, J.M. (2013). An effective implicit relevance feedback technique using affective, physiological and behavioural features. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 133–142). New York, NY: ACM. https://doi.org/10.1145/2484028.2484074

Moshfeghi, Y., Pinto, L.R., Pollick, F.E., & Jose, J.M. (2013). Understanding relevance: An fMRI study. In European Conference on Information Retrieval (pp. 14–25). Berlin, Heidelberg: Springer.

Mostafa, J., & Gwizdka, J. (2016). Deepening the role of the user: Neurophysiological evidence as a basis for studying and improving search. In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 63–70). New York, USA; ACM.

Nijholt, A., Tan, D., Pfurtscheller, G., Brunner, C., Millán, J.d.R., Allison, B., … Müller, K.R. (2008). Brain-computer interfacing for intelligent systems. IEEE Intelligent Systems, 23(3), 72–79. https://doi.org/10.1109/MIS.2008.41

Oliveira, F.T., Aula, A., & Russell, D.M. (2009). Discriminating the relevance of web search results with measures of pupil size. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2209–2212). New York, NY: ACM. https://doi.org/10.1145/1518701.1519038

Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., & Kaski, S. (2005). Combining eye movements and collaborative filtering for proactive information retrieval. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 146–153). New York, USA; ACM.

Ruotsalo, T., Jacucci, G., Myllymäki, P., & Kaski, S. (2015). Interactive intent modeling: Information discovery beyond search. Communications of the ACM, 58(1), 86–92.

Ruotsalo, T., Peltonen, J., Eugster, M., Głowacka, D., Konyushkova, K., Athukorala, K., et al. (2013). Directing exploratory search with interactive intent modeling. In Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (pp. 1759–1764). New York, USA; ACM.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology, 4(1). Article 32. https://doi.org/10.2202/1544-6115.1175

Treder, M.S., Schmidt, N.M., & Blankertz, B. (2011). Gaze-independent brain–computer interfaces based on covert attention and feature attention. Journal of Neural Engineering, 8(6), 066003.

Wenzel, M.A., Bogojeski, M., & Blankertz, B. (2017). Real-time inference of word relevance from electroencephalogram and eye gaze. Journal of Neural Engineering, 14(5), 056007. https://doi.org/10.1088/1741-2552/aa7590

Wenzel, M.A., Golenia, J.-E., & Blankertz, B. (2016). Classification of eye fixation related potentials for variable stimulus saliency. Frontiers in Neuroscience, 10, 23. Retrieved from https://www.frontiersin.org/article/10.3389/fnins.2016.00023. https://doi.org/10.3389/fnins.2016.00023

Zhu, X., & Davidson, I. (2007). Knowledge discovery and data mining: Challenges and realities. Hershey, PA: IGI Global.