Data Min Knowl Disc
DOI 10.1007/s10618-014-0379-5

A relative similarity based method for interactive patient risk prediction

Buyue Qian · Xiang Wang · Nan Cao · Hongfei Li · Yu-Gang Jiang

Received: 1 April 2014 / Accepted: 20 August 2014

Abstract This paper investigates the patient risk prediction problem in the context of active learning with relative similarities. Active learning has been extensively studied and successfully applied to solve real problems. The typical setting of active learning methods is to query absolute questions. In a medical application where the goal is to predict the risk of patients on certain disease using Electronic Health Records (EHR), the absolute questions take the form of "Will this patient suffer from Alzheimer's later in his/her life?", or "Are these two patients similar or not?". Due to the excessive requirements of domain knowledge, such absolute questions are usually difficult to answer, even for experienced medical experts. In addition, the performance of absolute question focused active learning methods is less stable, since incorrect answers often occur which can be detrimental to the risk prediction model. In this paper, alternatively, we focus on designing *relative* questions that can be easily answered by domain experts. The proposed relative queries take the form of "Is patient A or patient B more

Responsible editors: Fei Wang, Gregor Stiglic, Ian Davidson and Zoran Obradovic.

B. Qian (✉) · X. Wang · N. Cao (✉) · H. Li
IBM T. J. Watson Research, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA
e-mail: bqian@us.ibm.com

X. Wang
e-mail: wangxi@us.ibm.com

N. Cao
e-mail: nancao@us.ibm.com

H. Li
e-mail: liho@us.ibm.com

Y.-G. Jiang
Fudan University, 825 Zhangheng Road, Shanghai 201203, China
e-mail: ygj@fudan.edu.cn

Published online: 09 September 2014

similar to patient C?", which can be answered by medical experts with more confidence. These questions poll relative information as opposed to absolute information, and even can be answered by non-experts in some cases. In this paper we propose an interactive patient risk prediction method, which actively queries medical experts with the *relative similarity* of patients. We explore our method on both benchmark and real clinic datasets, and make several interesting discoveries including that querying relative similarities is effective in patient risk prediction, and sometimes can even yield better prediction accuracy than asking for absolute questions.

Keywords Patient risk prediction · Patient similarity · Active learning · Relative query · Reconstruction error · Counting set cover

1 Introduction

1.1 Background and motivation

The key idea of active learning is that a machine learning algorithm can achieve higher accuracy with fewer training labels if it is allowed to choose the data from which it learns. Active learning extends machine learning by allowing learning algorithms to typically query the labels from an oracle for currently unlabeled instances. Though enormous progress has been made in the active learning field in recent years (Zhuang et al. 2012), traditional active learning assumes that the questions prompted by a machine can be confidently answered by human experts, which may not be the case in many real world applications. This is particularly true in the domains that require proficient expertise to provide labels, such as medical informatics. The purpose of this work, Active patient Risk Prediction (ARP), is to explore easier active learning for medical data, so as to address the dilemma when doctors cannot confidently provide absolute labels.

Asking easier questions can effectively reduce the time and cost when applying active learning techniques to real world problems. For example, in medical informatics, patient similarity evaluation is an enabling technique for many research studies, such as risk stratification (i.e., grouping patients according to their disease condition risk), comparative effectiveness research, and predictive modeling. However, even for an experienced doctor, it is generally difficult to provide absolute labels in medical scenarios. For instance, it is difficult to judge whether two patients are exactly similar or not, since patients with the same disease may suffer from different comorbidities; but it is much easier to ask for relative comparison questions, such as *Is patient B more similar to patient A than patient C?*

To illustrate the advantage of using relative questions in medical applications, let us consider a simple example in a patient risk prediction problem as shown in Fig. 1. Consider the setting in Fig. 1a, where the query is an absolute question, it is difficult to tell whether the patient will suffer from the Alzheimer's later in his/her life based the MRI (Magnetic resonance imaging) scan, since the Alzheimer's cannot be predicted reliably even for experienced brain specialists. Consider the setting in Fig. 1b, which is another type of absolute question, it is also difficult to provide a binary answer that the two brains are similar or not, since similarity itself is a relative concept which may

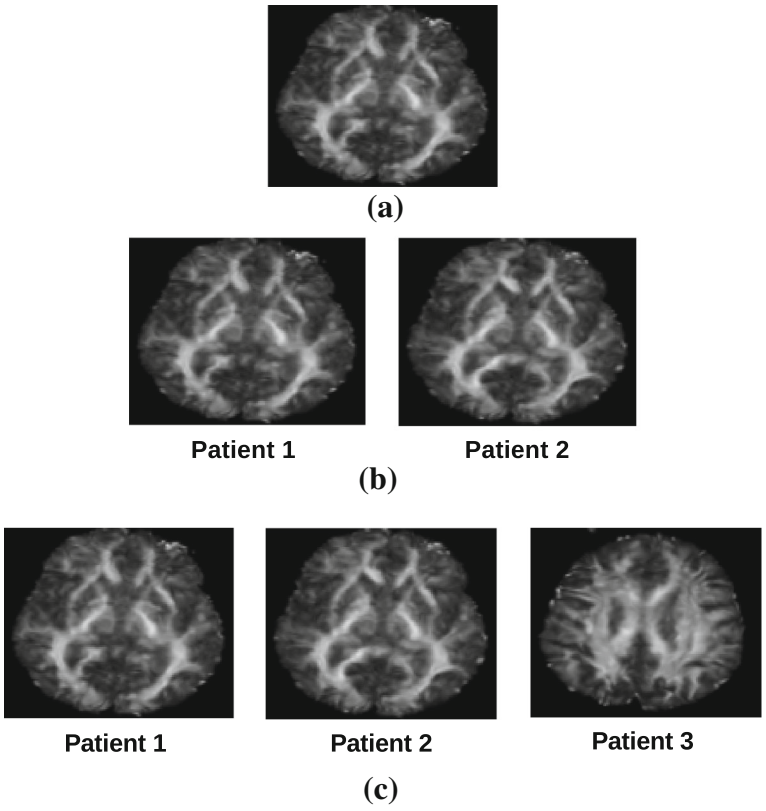

Fig. 1 Examples of absolute and relative queries on MRI images. **a** An absolute query: will this patient suffer from Alzheimer's later in his/her life? **b** An absolute query: are Patient 1 and Patient 2 similar or not? **c** A relative query: is Patient 2 or Patient 3 more similar to Patient 1?

not be answered independently with others. Now let us consider the setting in Fig. 1c, which shows an example of relative questions, it is significantly easier to provide a relative similarity amongst the three MRI scans, even a non-expert can answer the question by visually examining the three scans. The purpose of this work is to explore patient risk prediction in such context, i.e., *when absolute questions are difficult to query and obtain*. We propose to query relative information, rather than asking for exact labels, to better understand the neighborhood structure of instances.

1.2 Problem setting

In this work we explore active learning with relative questions in the context of the popular label propagation type of algorithms, such as `GFHF` (Zhu et al. 2003a) and `LGC` (Zhou et al. 2003). In this class of algorithms each instance (patient) is viewed as a node on a graph and has a weighted neighbor set (containing similar patients) which are collectively used to propagate labels from labeled instances to the unlabeled ones.

Therefore, the neighborhood structure (which instances are neighbors of each other and how to evaluate the similarity amongst themselves) is important to the performance of these algorithms since this determines where the labels are propagated to and how labels are propagated. In our formulation the neighborhood structure is learned by minimizing the reconstruction error of writing a data point as a weighted linear combination of its nearest neighbors. The given labels are then propagated to the unlabeled locations which are then further propagated and so on for an infinite number of steps.

1.3 Proposal

A key step of active patient risk prediction is to select informative queries, by answering which the prediction model is improved maximally. Our query selection strategy is, rather than asking for absolute questions such as labels, to ask humans to place ordering (possibly partial ordering) on the relative similarity of the neighbors to the instance that they are neighbors of. Specifically, a machine prompts a patient along with the patients who are similar to his/her (nearest neighbors), and then a medical expert is asked to sort or partially sort (from the most similar one to the least similar one) the neighboring patients according to the relative similarity to the patient that they are neighbors of. Since our active learning scheme is performed on neighbor sets rather than instances, we focus on selecting the most informative neighbor sets which we cast as a counting set cover problem. Counting set cover is an efficient combinatorial algorithm to perform entity ranking/selection, by using which we aim to locate the instance whose neighborhood is most influential to the graph structure. The proposed query scheme selects the most informative neighborhood to query, and the advice from human experts are enforced as constraints in the subsequent updating of the neighborhood structure, which are later used to help better propagate labels on the graph with the process being repeated. It is important to note that in our method there is no additional absolute labels that are added to the training data, rather the neighborhood weights are better estimated under the guidance from human experts.

The main advantage of querying neighborhood structure/weights is that the relative questions are easier to answer for medical experts. This is particularly useful to active learning in many specialized domains where the absolute questions are difficult to answer even for people with proficient domain knowledge, such as predicting a person's mental health condition based on the brain MRI scans, and inferring the patients' diseases based on the EHRs, which are the two real applications we shall focus on in this paper. The proposed method is appropriate to any data that can be cognitively understood by humans, such as MRI images which are visible to humans, and EHRs which are interpretable to humans. The majority of real data fits into this category, therefore, our method is applicable to most active learning applications. The proposed algorithm is computationally efficient and can be easily parallelized as discussed later. The promising experiment result demonstrates the effectiveness of the proposed approach, and validates our idea of querying relative similarities (neighborhood structure). In addition, our result indicates that sometimes querying neighborhood orderings can even achieve higher learning accuracy than using the same number of queries of labels.

1.4 Contribution

Our work makes the following technical contributions.

- We investigate a new form of knowledge injection (relative similarities rather than absolute labels) to active patient risk prediction.
- The proposed method can query both labeled and unlabeled patients.
- Our method is scalable to large medical problems since it divides a problem into a series of small problems each of which can be solved independently.
- We empirically show that the relative similarity can be a legitimate knowledge source (instead of absolute similarity or labels) in our method.

The rest of this paper are organized as follows. We begin our paper with a discussion of related studies in Sect. 2. We then present the details of the proposed method in Sect. 3, including the risk prediction model, problem formulations of the constrained neighborhood structure learning, query (neighborhood) selection strategy, and implementation issues. Our experiment in Sect. 4 explores the proposed method on a few benchmark datasets and real problems. We finally conclude our work in Sect. 5.

2 Related work

Our work is is broadly related to active learning and its applications, we shall now briefly review some previous studies. According to a recent survey on active learning (Settles 2009), existing active learning algorithms can be summarized into six categories based on the objective of the query selection.

Uncertainty sampling queries the instance about whose label the learning model is least confident (Lewis and Gale 1994; Culotta and McCallum 2005; Qian et al. 2013b) while *Query by committee* queries the instance about whose label the committee members (classifiers) most disagree (Muslea et al. 2000; Melville and Mooney 2004). The *Expected model change* query focus is on the instance that would impart the greatest change to the learning model (Settles et al. 2008). The *Expected risk Reduction* approach queries the instance which would minimize the expected future classification risk (Roy and Mccallum 2001; Guo and Greiner 2007; Kapoor et al. 2007) whilst the *Variance Reduction* query strategy chooses the instance which would minimize the output variance such that the future generalization error can be minimized (Zhang and Oles 2000). Finally *Density weighted method* query the instance which is not only uncertain but also representative of the underlying distribution of data (Settles and Craven 2008; Qian et al. 2013a; Cebron and Berthold 2009). However, all of them are label focused and hence are not directly comparable to our work. Existing approaches of active learning only focus on one aspect of active learning—the query strategy, and the other aspect of active learning—the design of questions—is not addressed.

A new direction in active learning is batch mode active learning (Hoi et al. 2006; Chattopadhyay et al. 2012) which asks the oracle a set of labels instead of a single label at a time. Although this is a more efficient querying method, it still requires the human experts to provide labels of a batch of instances and does not make the question itself mode efficient or easier. A novel direction proposed by Rashidi and Cook (2011) is a method that aggregates multiple instances into a generic active learning query based on

rule induction, and has been empirically demonstrated to perform more effective and efficient than querying labels. However, since it is a rule-based learning algorithm, its usefulness is limited to the cases that the data is represented in a low dimensional space and every feature has to be interpretable. Additionally, though a generic question it is still an absolute question as it requires human experts to have even stronger background knowledge than just querying labels. In contrast, we focused on designing a relative active learning query which could be answered by people without domain knowledge.

The only active learning work to our knowledge that queries pairwise relations is in the spectral clustering literature (Wauthier et al. 2012). This work presents an active spectral clustering algorithm that queries pairwise similarity, our work differs to this not only in learning setting (semi-supervised versus unsupervised) but also since we do not require the users to provide a real-valued pairwise similarity as they do rather just some orderings of the instances in a neighborhood set. Lately, researchers have been looking at applying data mining techniques to medical applications, such as patient similarity learning (Sun et al. 2012; Wang et al. 2012), patient risk prediction (Zhou et al. 2013; Davis et al. 2010), and patient pattern discovery (Norén et al. 2010; Wang et al. 2013). In addition, active learning has been successfully applied to EHRs data by Chen et al. (2013), in which there is an underlying assumption that human experts (doctors) can confidently answer the questions raised by machines. However, this may not be the case in many applications such as patient risk prediction where the target value (risk) is too difficult to query of. Ipeirotis et al. (2014) tried to address the noisy labeling issue in active learning by tolerating mistakes/errors provided by humans, but it does not cover the scenarios where the questions prompted by a machine is too difficult to answer even for a human with proficient domain knowledge. To address this scenario, our work focuses on developing easier active learning for patient risk prediction, where the major objective is to prompt answerable questions for human experts.

3 The ARP method

We in this section outline the proposed method of active patient risk prediction starting from describing the problem settings and notations. We then later in this section propose the risk prediction model with pairwise constraints (to enforce relative similarities), and the query selection strategy to identify the informative groups of patients to query of.

3.1 Preliminaries

3.1.1 Overview

Figure 2 shows the basic components and work flow of our proposed approach, which we refer to in this paper as Active patient Risk Prediction (ARP). In the first step, the ARP model takes the EHR data (possibly other types of medical data), and learns a prediction model, which will be later used to estimate the risk of a patient that will suffer from a particular disease. Then, the query selection strategy identifies an informative

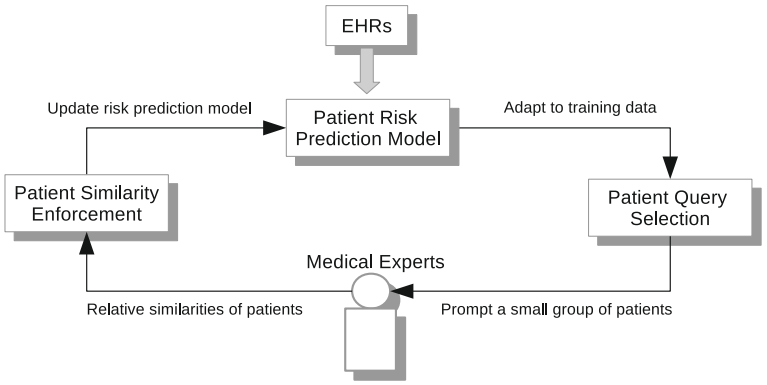

Fig. 2 The cycle of active patient risk prediction (ARP) model

group of patients (a set of neighboring patients) by solving a weighted counting set cover problem. In the next step of the ARP cycle, the medical experts are asked to provide the relative similarities (an ordering or partial ordering) to the identified group of patients based on the similarity to the patient that they are neighbors of. The provided relative similarity (neighborhood ordering) information will be enforced as pairwise constraints to update the risk prediction model. The process repeats until a desirable prediction accuracy is achieved.

3.1.2 Problem settings and notations

Formally, the problem being addressed in this work is described as follows. Given a set of n patients (instances) $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, we define a classification problem on a set of m possible diseases (labels). A small portion of the n patients were periodically examined, who are viewed as labeled instances, since we know what diseases they eventually have. Let a binary matrix B ($B \in \mathbb{R}^{n \times m}$) carry the given labels (diseases), where $b_{ij} = 1$ iff patient p_i suffers from disease j, and $b_{ij} = 0$ otherwise. Let $\mathcal{G}_{p_i}$ denote a small group of patients (neighbor set) which consists of the patients that are similar (nearest neighbors) to patient p_i. Note that for different patients the size of the neighbor set may differ. We then construct a graph of patients, which is fully defined by a patient similarity matrix S (sparse), where an entry $S_{ij} = 1$ if patient i and j are exactly the same, and $S_{ij} = 0$ if patient i and j are not similar. The relative similarities provided by medical experts are enforced to the learning of the patient similarity matrix S, such that a better patient similarity can be learned from human feedback, which in turns would produce a better estimate of patient risks. Note that we in our study do not require the notion of similarity defined in S to be symmetric, i.e., $S_{ij} \neq S_{ji}$ is possible and allowed. This paper will often refer to row or column vectors of matrices, for instance, i-th row and j-th column vectors of the matrix S are denoted as $S_{i\cdot}$ and $S_{\cdot j}$, respectively. The proposed ARP approach iterates between the learning of the patient similarity matrix S—updating a row of S in each iteration based on the relative similarities provided by medical experts, and the query selection—identify the most informative group of neighboring patients by ordering which the accuracy

of estimated risks would be significantly improved. In particular, our risk prediction model is built upon the linear neighborhood propagation algorithm (LNP) proposed by Wang and Zhang (2006), which learns the graph similarity matrix S by solving the reconstruction error as a quadratic program (QP) to guarantee the non-negativity of similarity. We shall next briefly review the LNP method, and then propose our pairwise constraints to incorporate the human feedbacks.

3.2 Background—linear neighborhood propagation

3.2.1 Learning of patient similarity

To present our work in a clear context, we now briefly review the reconstruction error and the Linear Neighborhood Propagation (LNP) framework, based on which we shall later in this section present the ARP approach. As introduced by Roweis and Saul (2000), the reconstruction error is defined as:

$$\mathcal{Q}(S) = \sum_{i=1}^{n} \| p_i - \sum_{p_j \in \mathcal{G}_{p_i}} S_{ij} p_j \|^2 \tag{1}$$

The reconstruction weight (patient similarity matrix) S is typically solved as a constrained least square problem or a linear system of equations, however, Wang and Zhang (2006) have shown that it also can be solved as a quadratic program (QP). The advantage of using a QP formulation is that additional constraints (such as non-negativity) can be added in, and thereby makes the formulation more flexible. Let $\mathcal{L}^i$ denote the local covariance matrix of patient p_i (the term "local" refers to the fact that the patient is used as the mean in the calculation of covariance), formally the definition of $\mathcal{L}^i$ can be expressed as $\mathcal{L}^i = (\mathbf{1}p_i - \mathcal{G}_{p_i})(\mathbf{1}p_i - \mathcal{G}_{p_i})^T$, where $\mathbf{1}$ denotes a column vector consisting of ones. Using the local covariance matrix, the reconstruction error problem can be reformulated to a series of small QP problems (one for each patient), since each row of S is independent of every other. Formally, a row vector $S_{i\cdot}$ (the weights used to reconstruct patient p_i using its neighbors) in the similarity matrix S can be solved as a QP problem as follows:

$$\begin{aligned} \min_{S_{i\cdot}} \quad & S_{i\cdot} \mathcal{L}^i S_{i\cdot}^T \\ s.t. \quad & S_{i\cdot} \mathbf{1} = 1; \\ & S_{ij} \geqslant 0, \ \forall \ j \in \{1, 2, \ldots, n\}. \end{aligned} \tag{2}$$

3.3 Learning of patient similarity with relative constraints

Instead of querying absolute questions, such as "will this patient suffer from a disease?" or "are the two patients similar or not?", our approach queries the relative similarities (neighborhood structure) amongst patients to improve the graph structure,

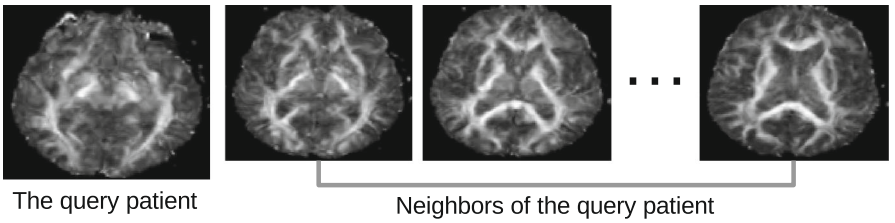

Fig. 3 An examplar query to doctors on strutural MRI data: to (partially) order the neighboring patients based on their realtive similarities to the query patient

which in turns will improve the performance of patient risk prediction. In each querying iteration, a doctor will be asked to order the neighbors of an patient in descending order based on their similarities. It is possible that sometimes a doctor may not be able to provide a complete ordering to a neighborhood, if this is the case the doctor can only provide a partial ordering, which still helps to constrain the learning of patient similarities. An exemplar relative query on structural MRI data is shown in Fig. 3. We see that even people without medical background can sometimes provide neighborhood orderings based on the visually observed similarities between patients. This implies that the benefit of the proposed ARP approach comes from that humans' visual perception being able to better understand the content of objects than can be calculated from the data.

The "unconstrained" version of patient similarity can be learned using Eq. (2). We now show how to encode the relative similarities (neighborhood ordering) to the formulation, such that the feedback from humans can be incorporated to the similarity matrix of patients. The QP formulation of the reconstruction error in Eq. (2) allows us to encode the neighborhood orderings as a set of linear constraints to the patient similarity matrix S. Let us take a simplified example to show the enforcement of relative similarities on a group of just three patients (the minimum number of patients for a relative similarity): given that a patient p_i has two neighboring (similar) patients p_a and p_b, and patient p_a is more similar to patient p_i than p_b is; we then can claim that the weight of p_a used to reconstruct p_i should be greater than that of p_b, i.e., we have $S_{ia} \geqslant S_{ib}$. This relative similarity can be encoded to the QP formulation using the linear constraint as shown below.

$$\begin{aligned} \min_{S_{i\cdot}} \quad & S_{i\cdot}\mathcal{L}^i S_{i\cdot}^T \\ \text{s.t.} \quad & S_{i\cdot}(\mathbf{e}^a - \mathbf{e}^b) \geqslant 0; \\ & S_{i\cdot}\mathbf{1} = 1; \\ & S_{ij} \geqslant 0, \ \forall\, j \in \{1, 2, \ldots, n\}. \end{aligned} \tag{3}$$

where $\mathbf{e}^a$ is a single-entry column vector with the a-th entry being one and all other entries being zeros. With the transitivity of inequality, we can enforce a complete ordering on a group of neighboring patients using a set of concatenating constraints. For example, if we require that patient p_a, p_b and p_c are decreasingly similar to patient p_i, i.e., in terms of patient similarity we have $S_{ia} \geqslant S_{ib} \geqslant S_{ic}$. This ordering can be

enforced using two concatenating constraints, i.e., $S_i.(\mathbf{e}^a - \mathbf{e}^b) \geqslant 0$ and $S_i.(\mathbf{e}^b - \mathbf{e}^c) \geqslant 0$.

3.4 Query selection of patient neighborhood

With the QP formulation shown in Eq. (3), we can learn the patient similarity matrix S under the guidance of the relative similarities provided by medical experts. Then the next question to address in our study is to select informative patient neighborhoods, which will be prompted to human experts as queries.

We begin this section with the overview of the intuition behind the proposed method and then providing full details for the reproducibility of our experimental results. Our work deviates from existing active learning methods by querying not the risk of a patient or the absolute similarity between two patients, rather querying the neighborhood structure of patients. Hence, our query strategy aims to choose the patient neighborhood which if queried (sorted by human experts) will have the most significant impact in terms of better propagating the risk of diseases on the patient graph. Since each neighbor set of patients naturally forms a subset of the n patients, we propose to use counting set cover to estimate the importance of a patient neighborhood. Benefits of using counting set cover include (i) neighbor sets that are essential to maintain the graph structure can be naturally found through solving a set cover problem, and (ii) counting with different weighting schemes would emphasize different notions of the importance of graph structure, which enriches the flexibility of active neighborhood selection.

Before we present the details of counting set cover, we first briefly review the set cover problem and its efficient approximation. Recall a set cover problem consists of two parts: (1) a universe which in our case is the patient set $\mathcal{P}$ containing all the n patients, and (2) a set of subsets of $\mathcal{P}$ which in our case is the n patient neighbor sets that correspond to the n patients, i.e., $\mathcal{G} = \{\mathcal{G}_{p_1}, \mathcal{G}_{p_2}, \ldots, \mathcal{G}_{p_n}\}$. We say a subset $\mathcal{G}_\mathcal{S}$ of $\mathcal{G}$ ($\mathcal{G}_\mathcal{S} \subset \mathcal{G}$) is a cover of the universe $\mathcal{P}$ if every patient in $\mathcal{P}$ appears at least once in $\mathcal{G}_\mathcal{S}$. In other words, the union of the subsets in $\mathcal{G}_\mathcal{S}$ is the universe, i.e., $\cup\, \mathcal{G}_{\mathcal{S}_i} = \mathcal{P}$. A cover $\mathcal{G}_\mathcal{S}$ that has minimum (possibly weighted) cardinality is called a minimum set cover. The set cover problem, which aims to identify such a minimum cover, can be formulated as the following integer program.

$$\begin{aligned} \min \ & \sum_{i=1}^{n} \mathcal{C}_i \\ s.t. \ & \sum_{j=1}^{n} \beta_{ij}\mathcal{C}_j \geq 1, \qquad \forall\ p_i \in \mathcal{P} \\ & \mathcal{C}_i \in \{0, 1\}. \end{aligned} \tag{4}$$

where $\mathcal{C}_i$ is the indicator of the subset $\mathcal{G}_{p_i}$ (the neighborhood of patient p_i), which is set to 1 iff the neighborhood $\mathcal{G}_{p_i}$ is part of the minimum set cover, and 0 otherwise. β_{ij} indicates whether the patient p_i exists in the neighborhood $\mathcal{G}_{p_j}$ (the neighborhood

of patient p_j), i.e., $\beta_{ij} = 1$ if $p_i \in \mathcal{G}_{p_j}$ and $\beta_{ij} = 0$ otherwise. The first constraint in Eq. (4) is to guarantee that every patient in $\mathcal{P}$ is covered at least once in the solution, and the second constraint is to enforce the set cover indicator to be binary, i.e., $\mathcal{C}_i$ is either 0 or 1.

The set cover is a well studied NP-hard problem. In our work we adopt the following greedy approximation algorithm to solve the set cover problem: in each step, choose the subset $\mathcal{G}_{p_i}$ that contains the most uncovered patients; Repeat this process until all patients are covered. This simple greedy approach finds a set cover with at most $c^* log_e n$ sets, where an optimal solution contained c^* sets. With random initializations this method produces multiple close to minimum set covers. This allows us to count the number of solutions that each patient neighborhood $\mathcal{G}_{p_i}$ participates in, and make the estimation of each neighborhood's significance easier. Let $\mathcal{Q}(\mathcal{G}_{p_i})$ denote the significance of patient p_i's neighborhood to maintain the graph structure. Formally, we can write the $\mathcal{Q}(\mathcal{G}_{p_i})$ as a weighted counting set cover problem shown as follows:

$$\mathcal{Q}(\mathcal{G}_{p_i}) = \sum_{z_j \in \mathcal{Z}} \gamma(\mathcal{G}_{p_i}, z_j) w(\mathcal{G}_{p_i}) \tag{5}$$

where z_j denotes a close to minimum set cover solution, and $\mathcal{Z}$ denotes the collection of the multiple close to minimum set covers. $\gamma(\mathcal{G}_{p_i}, z_j)$ indicates whether the neighborhood $\mathcal{G}_{p_i}$ is part of the (close to minimum) set cover z_j. Formally, $\gamma(\mathcal{G}_{p_i}, z_j) = 1$ if $\mathcal{G}_{p_i} \in z_j$, and $\gamma(\mathcal{G}_{p_i}, z_j) = 0$ otherwise. $w(\mathcal{G}_{p_i})$ is the counting weight of patient neighborhood $\mathcal{G}_{p_i}$, which also can be viewed as defining the querying preference of patient neighborhoods. We in our study consider the following two weighting schemes:

- Uniform: $w(\mathcal{G}_{p_i}) = 1, \forall\ \mathcal{G}_{p_i} \in \mathcal{G}$. A baseline weighting scheme that assigns a uniform weight to all patient neighborhoods. The underlying assumption of this weighing scheme is that all patient neighborhoods are equally important in the counting, therefore, from the weighting perspective they are equal likely to be selected.
- Connectivity: $w(\mathcal{G}_{p_i}) = \sum_{j=1}^{n} S_{ji}$. In plain words, the counting weight of a neighborhood $\mathcal{G}_{p_i}$ is proportional to the frequency that the members of $\mathcal{G}_{p_i}$ are used to reconstruct others. This weighting scheme is node (patient) connectivity based, which assigns higher weights to the patient neighborhoods that are located in the "dense" area of the patient similarity matrix S. That is where the learning algorithm is more liked to be confused. This implies that the weighting scheme prefers to query the patient neighborhoods that are highly connected to others, including both within and outside the neighborhood, since they are more influential in maintaining the key structure of the patient graph.

Once an informative patient neighborhood is identified by the counting set cover strategy, a medical expert will be asked to sort the neighbors in descending order with respect to the similarity to the query patient that they are neighbors of. The relative similarities obtained from human experts will be later incorporated to the learning of the patient similarity matrix S, which in turns would improve the prediction accuracy of patient risks. This process repeats until a desirable prediction accuracy is achieved,

or some stopping criterion is satisfied. Note that since the learning of one row $S_{i\cdot}$ in the matrix S is independent of other rows, each feedback (query) from human experts would only change a single row in S. This shows another advantage of the proposed method on active patient risk prediction: there is no need to updated the entire model in each querying iteration.

3.5 Risk estimation

After the active learning of patient similarity matrix S, we can perform patient risk prediction using graph diffusion methods. In particular, we propagate the given risks of diseases on the patient graph using the similarity matrix S. In our method, each row of the patient similarity matrix S sums to one, thereby S can be readily used as the transition matrix and perform a random walk on the graph to infer patient risks. In each risk propagation iteration, the state (i.e. the risks of diseases) of each patient is partially (with a rate of λ) adjusted by risk values that flow on the graph, but still preserves a portion (with a rate of $1-\lambda$) of the given true risks. Let R denote the predicted patient risk and B denote the given true risk, the state (risk) of patients at time $t+1$ can be inferred from the previous state of patients at time t.

$$R^{t+1} = \lambda S R^t + (1-\lambda)B \tag{6}$$

Let R^{∞} denote the patient state (risk) after infinite random walk steps, the state of patients eventually converges to a steady-state probability as follows.

$$R^{\infty} = (1-\lambda)(I-\lambda S)^{-1}B \tag{7}$$

where I denote the identity matrix. Note that the value of $(1-\lambda)$ can be interpreted as the probability of the restart (jump back to the initial state) in a random walk. The restart is a necessary step in the risk propagation process, otherwise the problem would reduce to a global solution, which is equivalent to the result of PageRank. $(1-\lambda)$ also can be viewed as the constant to penalize the changes to the initial risks. In practice, if the given risk B is relatively more complete (B is a dense matrix), we should set a smaller value for λ. On the contrary, the value of λ should be larger if the given risk B is relatively less complete (B is a sparse matrix) or contains considerable amount of noise.

3.6 Implementation

The proposed method mainly aims to perform effective patient risk prediction by asking medical experts answerable or easier questions. The ARP algorithm is summarized in Table 1. Since our method may query on both patients with given risks and patients without given risks, it is possible that our approach may choose to query the same patient neighborhood repeatedly. In practice, this issue can be solved by simply enforcing that each patient neighborhood only can be queried once. If the sparsity is introduced to the patient similarity matrix S, such as using k-nearest neighbor,

Table 1 The ARP Algorithm

Input:

Patients $\mathcal{P} = (p_1, p_2, \cdots, p_n)$, initial patient risk B, the parameter λ

Training Stage:

```
do {
    (1) solve patient similarity with relative guidance using Eq.(3);
    (2) find the next patient neighborhood to query using Eq.(5);
    } until desirable prediction accuracy is achieved;
perform risk propagation on the patient similarity graph using Eq.(7);
```

Output:

The patient similarity S and the patient risk prediction R

the proposed algorithm would be computationally efficient and applicable to large scale problems, since (i) the QP learning of patient similarity can be efficiently solved and (ii) only a single row in S needs to be updated after each feedback from human experts.

In order to reduce the human efforts of ordering a patient neighborhood and avoid excessive patient risk propagation, the number of neighbors of a patient needs to be limited to a small value. In our implementation, we discard the neighbors whose weights in $S_{i\cdot}$ are under a certain threshold (0.01 in our experiment). The QP formulation defined in Eq. (2) is a standard problem, thereby can be solved using any standard QP solvers. We in our experiment used the build-in QP solver of Matlab.

In addition, the proposed ARP method can be easily scaled in a number of ways.

- One may employ kd-trees (Panigrahy 2008) or locality sensitive hashing (Gionis et al. 1999) techniques to efficiently construct the patient neighborhoods $\mathcal{G}$.
- The learning of patient similarity S defined in Eq. (2) can be easily parallelized, since the weights to reconstruct each patient using their neighbors are solved independently of the weights that are used to reconstruct other patients.
- The patient risk propagation defined in Eqs. (6) or (7) also can be parallelized as the matrix manipulation can be parallelized.
- To accelerate the selection of informative patient neighborhoods, one may use more efficient counting set cover algorithms such as compressed-IC (Gionis et al. 2012).

4 Empirical evaluation

We in this section attempt to understand the strengths and relative performance of the proposed approach `ARP`. As discussed in Sect. 3.4, there are two versions of `ARP` that are evaluated in our experiment: (i) `ARP-Uniform` which counts the set covers using the uniform weighting scheme, and (ii) `ARP-Connectivity` which counts

set covers using the connectivity weighting scheme. In particular we wish to answer how well our method compares to the following baseline methods:

1. `Random+Risk` A baseline with the setting as follows, (i) the patient similarity matrix S is learned using Eq. (3) with relative feedback, (ii) the selection of patient neighborhood to query is random, (iii) the patient risk is inferred using Eq. (7). This baseline has the same learning settings as our `ARP` method but selecting random neighborhood to query. We chose this as a baseline to demonstrate that the counting set cover scheme presented in Eq. (5) significantly outperforms the random selection of neighborhoods.
2. `Active Harmonic Function` A state-of-the-art active graph propagation method proposed by Zhu et al. (2003b). In our case this method queries human experts with the patient risks, not the relative similarity. We chose this as a baseline to evaluate how well our `ARP` method compares to traditional label focused active learning methods.

The surprising answer is that the proposed `ARP` method performs comparably as typical label focused active learning despite not adding more labels (patient risks in our experiment) rather just improving neighborhood structure. Given this, an interesting question is, "Is the good performance of `ARP` method due to the query selection scheme or the learning technique of patient similarity?". To investigate this we compare the `ARP` method with the following two additional baseline methods:

3. `Active+Risk` A baseline method that is exactly the same as the `ARP`, but querying human experts with the patient risks rather than the relative similarities of patients, i.e., asks for the risk of patient p_i rather than the order of $\mathcal{G}_{p_i}$. In this method, the set covers are counted using the connectivity weighting scheme. We chose this as a baseline to compare, under our query selection scheme, the strengths between risk values and relative similarities in active patient risk prediction.
4. `Random+Relative` A baseline methods that is exactly the same as the `ARP`, but randomly selecting patient neighborhood to query, i.e., replace the counting set cover scheme with a random selector. We chose this as a baseline to evaluate the performance of our patient similarity learning method on a bad (random) query selection scheme.

We compare the performance of the above six methods (including the four baseline methods and the two versions of `ARP`) on three sets of real data.

- *Benchmark data* Breast Cancer and Diabetes from UCI machine learning repository (Asuncion and Newman 2007), both of which are extensively tested and have shown to be useful in the evaluation of learning methods.
- *Structural MRI data* We evaluate our method on a collection of the brain structural MRI scans of over 1,000 real patients, who were routinely clinically examined for years so that we know whether they have the Alzheimer's disease eventually. The task here is to predict the risk of each patient that will suffer from the Alzheimer's disease later in their life.
- *EHR data* We used a EHR dataset extracted from a real-world warehouse which consists of the records of 319,650 patients over 4 years. We selected Congestive Heart Failure (CHF) patients as a study case and predicted their risk of CHF onset.

There are two main parameters used in the six methods, i.e. the σ in `Active Harmonic Function` and the λ in `ARP`, both of which are selected using cross-validation. To reduce the number of patients in a neighborhood, we in the `ARP` approach discard the neighbors with a weight less than 0.01, in the `Active Harmonic Function` use k-nearest neighbors method when constructing the patient similarity graph (k is usually 7–15 depending on the data).

Since the questions queried in our model is the relative similarity of patients rather than the risk/label, it is possible (but unlikely in practice) that a neighborhood ordering provided by a medical expert does not make changes to the patient similarity graph. In other words, it is possible that in an active learning iteration our `ARP` learning model gains nothing and remains the same. To evaluate this, we define a measure called *hit-rate*, which refers to the fraction of times that the queried neighborhood orderings changed the patient similarity graph. The *hit-rate* indicates the success of the relative queries in terms of changing the patient risk prediction model. In particular, a low *hit-rate* implies querying the relative similarity of patients did not bring much new information to the risk prediction model, while a 100 % *hit-rate* indicates that every queried patient neighborhood made some changes to the risk prediction model. There are two evaluation measures used in our experiment.

- *Error rate* the rate of mis-predicted patients, which measures the accuracy of patient risk prediction models.
- *Hit-rate* the fraction of times that the feedbacks from medical experts made changes to the risk prediction model, which measures the efficacy of the active scheme of our `ARP` method.

4.1 Experiment 1—Benchmark datasets

4.1.1 Dataset and experiment settings

We first evaluate the proposed `ARP` method on two benchmark medical datasets from UCI repository, i.e., Breast Cancer and Diabetes. The Breast Cancer dataset contains 569 patients that are represented using 30-dimensional data vectors, which are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The features describe the characteristics of the cell nuclei presented in the images. All the patients are categorized as either malignant (positive) or benign (negative). The Diabetes datasets consists of 768 patients, who are all females of Pima Indian heritage and at least 21 years old. Each patient is represented by a 8-dimensional data vector, and are categorized as either diabetic patient (positive) or normal patient (negative). In this experiment, the learning task is to predict the risk of a patient that suffers from breast cancer or diabetes. Performing active learning on the two data is difficult, since the labeling of patients based on the medical measurements requires extensive medical experience. However, it would be easier to ask for the ordering of patient in a neighborhood, which requires less medical knowledge compared to absolutely saying that a patient is normal or sick. For the reproducibility of our experimental result, we use the labels of patients to generate relative similarities. In particular, given a query patient and the corresponding neighborhood, we enforce the neighboring patients with

the same label as the query patient to be more similar to the query patient than the patients with a different label, but do not enforce ordering amongst the patients with the same label. In each active learning iteration, we only enforce a partial ordering to the learning of patient similarity, which is weaker than a complete ordering of neighborhood.

4.1.2 Results and discussion

For both Breast Cancer and Diabetes datasets, in each trial, we randomly and equally divide the patients into four groups. One group is used for training and the rest patients are used for test, and we do a rotation of the training group among the four groups of patients. For each training set, 30 additional labels or relative similarities (depends on the methods) of patients are added to the training set step by step, which are used to show the accuracy improvement of patient risk prediction along with the increase of active queries. The experiments are repeated for 100 times, and the mean error rates with the standard deviations are reported in Fig. 4, where the result on Breast Cancer dataset is shown in Fig. 4a and Diabetes dataset is shown in Fig. 4b.

In the result, we see that the two random querying methods `Random+Risk` and `Random+Relative` are not much helpful to the learning accuracy. Especially on the Breast Cancer dataset, the error rate keeps flat during the 30 queries of labels or relative similarities. This confirms the motivation and necessity of active learning scheme since asking randomly selected questions may not effectively improve the performance of patient risk prediction. Among the four active querying methods, the overall best accuracy is achieved by `Active Harmonic Function` and `Active+Risk` methods, which implies that querying patient labels provides stronger knowledge to the prediction model than the relative similarity does. However, it can be seen that our `ARP-Connectivity` methods achieves comparable (sometimes even better) prediction accuracy with the methods without adding more labels, which confirms our motivations that asking easier questions (relative similarities) can also

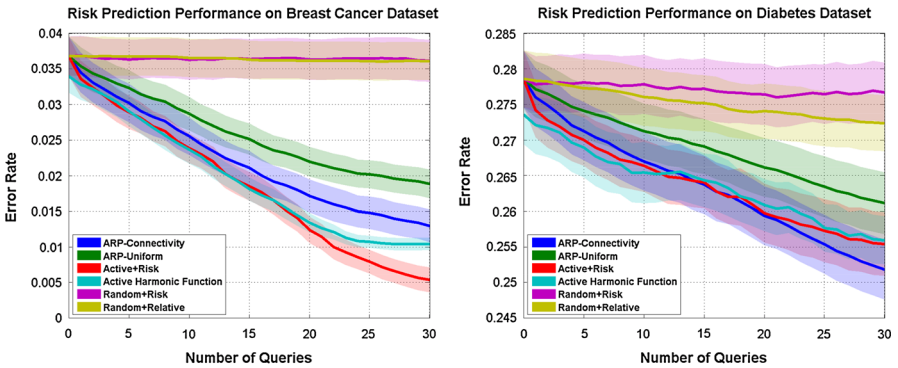

Fig. 4 Performance evaluation on UCI benchmarks (standard deviation is marked by *shade*). **a** Error rate on breast cancer, **b** error rate on diabetes

Table 2 Hit-rate comparison on the two UCI benchmarks

Dataset name	Hit-rate of ARP	Hit-rate of random query
Breast cancer	100 %	87.03 %
Diabetes	100 %	88.67 %

efficiently improve the performance of risk prediction. In addition, we can see in the result of Diabetes dataset that our `ARP-Connectivity` method sometimes outperforms the two active methods that query labels, and `Random+Relative` outperforms `Random+Risk`. This shows that the proposed `ARP` method is more noise resistant, since label is a stronger type of supervision than relative similarity, and thus the label of a noisy data vector would be more destructive to the prediction model. For the two weighting schemes used in counting set cover, the connectivity weight significantly outperforms the uniform one, since by using the former weight the graph structure is factored into the counting.

The hit-rate comparison on the two UCI benchmark datasets are reported in Table 2. We see that the hit-rate of the `ARP` method is always 100 %, which implies that the patient neighborhoods selected our `ARP` method are more likely to be improvable. Since the randomly selected neighborhoods are not noticeably helpful in the updating of patient similarity matrix S, the usefulness of our patient neighborhood selection scheme is validated.

4.2 Experiment 2—prediction of Alzheimer's disease

4.2.1 Dataset and experiment settings

The structural MRI dataset consists of 1,005 patients, and the raw scans were collected from real clinic cases. An example of the 3D MRI images used in our experiment is shown in Fig. 5a. This is a *new dataset* and will be made publicly available. There are two types of MRI scans that were collected from the subjects. (1) *FLAIR*: Fluid attenuated inversion recovery is a pulse sequence used in MRI, which carries the white matter hyper-intensity of a brain. (2) *GRAY*: Gray MRI images which only reveals the activities in the gray matter of a brain. Figure 5b shows the region of white matter. The MRI raw scans are in 3D, and each voxel has a value from 0 to 1, where 1 indicates that the structural integrity of the axon tracts at that location is perfect, while 0 implies either there are no axon tracts or the tracts are shot (not working). The raw scans are preprocessed (including normalization, denoising and alignment) and then restructured to 3D matrices with a size of $134 \times 102 \times 134$. To further reduce the dimension of data, we divide the brain into a set of biological regions (104 for white matter and 46 for gray matter, such as Hippocampal, Cerebellum, Brodmann, Fornix, $\cdots$), and take the mean value of each region to represent the entire region. This is a common data preprocessing procedure for brain images in neuroscience. The white and gray matter regions (marked by blue) are shown in Fig. 5b and c, respectively.

When the MRI scans of patients were collected, their cognitive function scores (including semantic, episodic, executive and spatial, which range between -2.8258 and 2.5123) were also periodically acquired using a cognitive functioning test. The

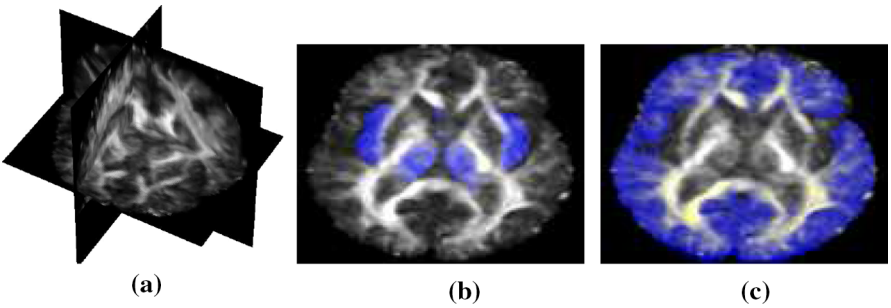

Fig. 5 Examples of structural MRI image. **a** 3D MRI scan, **b** white mater, **c** gray matter

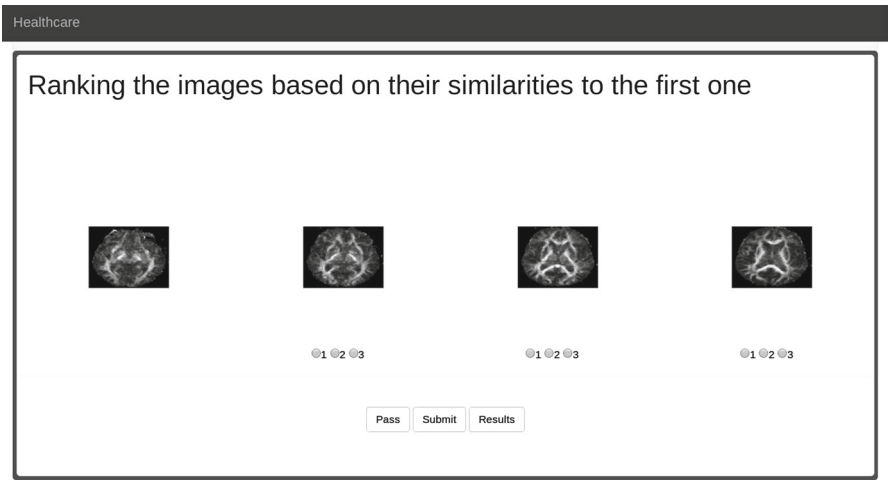

Fig. 6 The user interface for medical experts to input relative similarities

cognitive scores indicate the mental health condition of patients, and are used as the evidence to infer if a patient has the Alzheimer's disease or not. Since the 1,005 patients in the dataset were routinely periodically examined, we know whether they eventually suffer from Alzheimer's disease. This is used as the ground truth in our experiment. The risk prediction problem on this dataset is to determine if a patient will be normal, mildly cognitively impaired (MCI), or dementia later in their life based on their current brain images. Inferring the occurrence of Alzheimer's disease based on MRI scans is a difficult task, even for experienced brain specialists. However, the proposed `ARP` method asks for relative similarities amongst patients instead, which makes active learning on such dataset significantly easier. For example, though it is difficult to judge a person's mental health condition based on MRI images, doctors can confidently provide the relative similarities amongst patient by visually comparing the similarities and differences between the MRI images. To collect the relative similarities amongst patients, we create a web user interface for doctors, as shown in Fig. 6, to input their feedbacks. Note that the relative similarities of patients used in our experiment are provided by real brain specialists and neuroscientists.

4.2.2 Results and discussion

For both FLAIR and GRAY MRI images, we in each random trial equally divide the 1,005 patients into three groups. Patients in one group are used for the training, and the remaining patients are used in the test. The training group rotates until every group of patients being used as the training set once. To evaluate the performance improvement of patient risk prediction as the number of active queries increases, we gradually add 30 additional labels or relative similarities (depending on the method used) of patients to the training data. The experiments are repeated for 100 times. The mean error rates along with the standard deviations are reported in Fig. 7, where the result on Breast Cancer dataset is shown in Fig. 4a and Diabetes dataset is shown in Fig. 4b, where Fig. 7a shows the result on FLAIR type of MRI images and Fig. 7b shows the result on GRAY type of MRI images.

Observing the result, we see that the performance of the two random querying methods, i.e. `Random+Risk` and `Random+Relative`, are very weak as they do not improve the learning accuracy and sometimes are even destructive. From the results we can observe that the connectivity is definitely a better weighting scheme for our `ARP` approach compared to the uniform weighting scheme. It also can be seen that in this experiment there are two methods using our query selection scheme, `ARP-connectivity` and `Active+Risk`, both of which significantly outperform the `Active Harmonic Function`. This demonstrates the effectiveness of our counting set cover strategy. Surprisingly, the performance of `ARP-connectivity` is not just comparable to the label focused active learning methods, but even performs better. We investigate into this, and find this is caused by the complexity of the medical data. The MRI images are complex objects, therefore, for a risk prediction model it is difficult to understand the images directly from the features, which in turns make the correct construction of neighborhood structure difficult. Hence, in this application providing a few additional labeled patients may not improve the prediction model much, but providing a few key neighborhood orderings (relative similarities) would

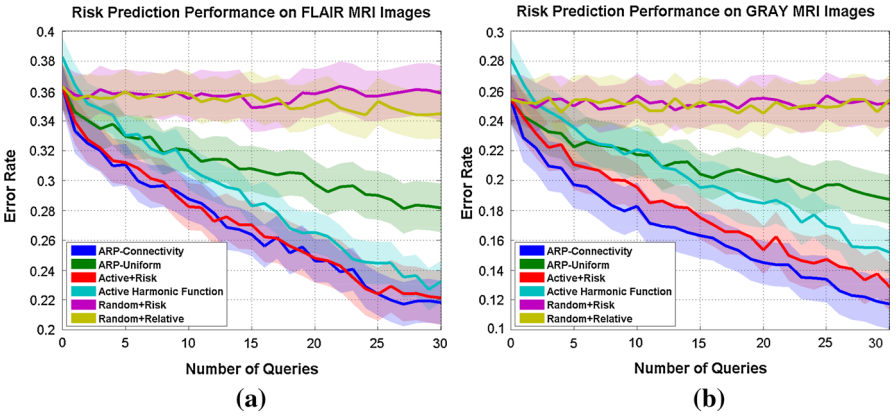

Fig. 7 Performance evaluation on MRI scans (standard deviation is marked by *shade*). **a** Error rate on FLAIR scans, **b** error rate on GRAY scans

Table 3 Hit-rate comparison on the two types of MRI images

Type of MRI image	Hit-rate of ARP	Hit-rate of random query
FLAIR	100 %	87.33 %
GRAY	100 %	86.60 %

significantly enhance the graph structure and better propagate the risks on the graph of patients. This confirms the advantage of querying relative similarities, and implies that the proposed `ARP` model is more suitable for the learning problems involving complicated objects or high dimensional data.

In Table 3, we see that for FLAIR type of MRI scans the hit-rate of the `ARP` method is 100 %, while the `Random+Relative` method only reaches a hit-rate of 87.33 %; for GRAY type of MRI scans the hit-rate of `ARP` is also 100 %, while the random selection being only 86.60 %. From the result, we see that same as our `ARP` model, the baseline method `Random+Relative` also makes changes to the risk prediction model by modifying the patient similarity matrix S. However, the changes of graph structure made by the `Random+Relative` method are not comparable to the changes made by our `ARP` method, and are not helpful to the risk prediction performance. This implies that the random query selection may not improve the learning performance, which confirms the demand for the proposed query selection scheme.

4.3 Experiment 3—prediction of congestive heart failure

4.3.1 Dataset and experiment settings

We used a real-world EHR (Electronic Health Record) data warehouse including the records of 319,650 patients over 4 years. We selected Congestive Heart Failure (CHF) patients as a study case and predicted their risk of CHF onset. We defined CHF diagnosis using the following criteria (Wu et al. 2010): (i) International Classification of Diseases - Version 9 (ICD-9) diagnosis codes of heart failure appeared in the EHR at least twice, indicating consistency in clinical assessment; (ii) at least one CHF-related drug prescription. The diagnosis date of a confirmed CHF patient was defined as the first occurrence of the heart failure related diagnosis code. With this criteria, we extracted from the database 1,127 CHF case patients. Following the case-control match strategy in (Wu et al. 2010), a primary care patient was eligible as a control patient if he did not meet the CHF diagnosis criteria and had the same PCP as the case patient. Approximately ten eligible clinic-, gender-, and age-matched (in 5-year age intervals) controls were selected for each heart failure case. In situations where ten matches were not available, all available matches were selected. Following this strategy, we obtained CHF 3,850 control patients, which means each case patient was matched with three controls on average.

For all case and control patients we extracted their Hierarchical Condition Categories (HCC) codes from the EHR database as medical features. HCC codes are higher-level categorization of patients' diagnosis and it is correlated to the more detailed ICD-9 codes. We only considered the medical records that occurred from

540 days prior to the diagnosis date till 180 days prior to the diagnosis date. In other words, we used about a year worth of data to make prediction at least half a year before the disease onset. Patients who had insufficient amount of records were not included. For control patients, we set the last day of their available records as the diagnosis date and followed the same rule. In total there were 186 unique HCC codes for all the 4,977 patients. The medical records were encoded in a binary $4,977 \times 186$ matrix. The (i, j)-th entry was set to 1 if HCC code j was observed on patient i, and 0 otherwise. In total 65,467 medical records were considered, which indicates our input data was extremely sparse (7% nonzero entries).

Predicting Congestive Heart Failure (CHF) of patients based on the HCC codes is difficult, however, the proposed `ARP` method makes active learning techniques applicable to such difficult learning tasks, since it asks for easier relative questions instead of the absolute questions used in typical active learning. For example, though it is difficult to tell if a patient will suffer from CHF or not, doctors can confidently provide the relative similarities of patients by comparing their HCC codes. To collect the relative similarities for our evaluation, we also create a web user interface for doctors, just like we did in the Alzheimer's Disease experiment. Note that the relative similarities of patients used in this experiment are provided by real experienced doctors.

4.3.2 Results and discussion

In this experiment, we in each random trial equally divide the nearly 5,000 patients into 3 groups. One of the three groups is for the training, while all the remaining patients are used for the testing. We then rotate the group being the training set until all groups of patients are covered. To simulate the scenarios that different numbers of active queries have been performed, we add 150 additional labeled patients or relative similarities (depends on the method) to the training data one-by-one. The experiment is repeated for 50 times. The mean error rates with their standard deviations are reported in Fig. 8, where the mean error rate is shown in bold curve and the standard deviation is presented by shaded borderline.

From the result shown in Fig. 8, we can sort the overall performance of the six methods in the descending order as follows: (1) `Active+Risk` $\succ$ (2) `Active Harmonic Function` $\succ$ (3) `ARP-Connectivity` $\succ$ (4) `ARP-Uniform` $\succ$ (5) `Random+Risk` $\succ$ (6) `Random+Relative`. From the observation we see that `Active+Risk` outperforms `ARP-Connectivity` and `Random+Risk` outperforms `Random+Relative`, which indicates that in risk prediction problems label is a stronger type of supervision than relative similarity, since the only difference in both the two pairs of methods is the supervision. It can be observed that the best performance is achieved by the two label focused active learning methods, i.e. `Active+Risk` and `Active Harmonic Function`, which is in line with our expectation. However, it is important to note that in this setting the two label focused methods assume that the medical experts being able to provide perfect labels of patients based on the HCC codes, which may not be realistic in practice. Thus their performance can be viewed as the upper bound in this experiment. We see that the two `ARP` variants, i.e. `ARP-Connectivity` and `ARP-Uniform`, achieve comparable performance to the two label focused methods, though there is no additional label added, but rather the

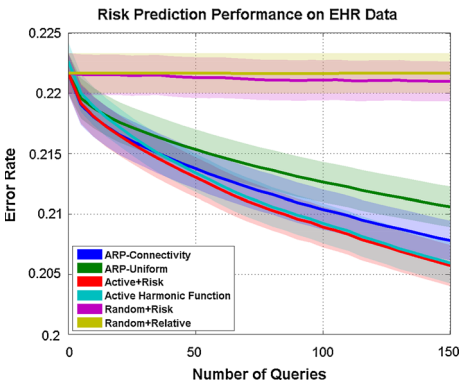

Fig. 8 Performance evaluation on EHR dataset (standard deviation is marked by *shade*)

relative similarities. This shown the effectiveness of our ARP method. In addition, we can see that the connectivity is a better weighting scheme for the counting set cover strategy. The result also shows that the methods using randomly selected queries do not noticeably improve the accuracy of patient risk prediction, though there are 150 labels or relative similarities added to the training data. This motivates the need for active query selection, since additional training data does not necessarily improve the prediction accuracy.

5 Conclusion

In this paper we present an interactive system to predict the patient risks of suffering from certain diseases, and propose an active learning method that queries the relative similarities of patients. The proposed relative queries take the form of, "Is patient i or patient j more similar to patient k?". This relative type of questions is easier being answered by medical experts, which in turns make active learning methods more applicable to difficult medical problems or the medical learning tasks involving complex features (such as MRI images). The proposed ARP method is easy to implement, and can be scaled to solve large healthcare learning problems using parallel/distributed computing. The surprising empirical results on real-world medical problems demonstrate the usefulness of our ARP method, as querying for the relative similarities of patients can achieve comparable and in some cases even better prediction performance than querying absolute questions on patients, while the latter type of questions is significantly more difficult to answer. It is important to note that the ARP method makes active learning on difficult medical problems possible, e.g. the prediction of Alzheimer's disease and congestive heart failure, where the traditional active learning approach cannot be applied. This is significant since active learning is especially needed in difficult medical learning tasks, and our work is a first step towards this goal.

References

Asuncion A, Newman D (2007) Uci machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html

Cebron N, Berthold MR (2009) Active learning for object classification: from exploration to exploitation. Data Min Knowl Discov 18(2):283–299

Chattopadhyay R, Wang Z, Fan W, Davidson I, Panchanathan S, Ye J (2012) Batch mode active sampling based on marginal probability distribution matching. In: KDD, pp 741–749

Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, Xu H (2013) Applying active learning to high-throughput phenotyping algorithms for electronic health records data. JAMIA 20:e253–e259

Culotta A, McCallum A (2005) Reducing labeling effort for structured prediction tasks. In: Proceedings of the 20th national conference on artificial intelligence—vol 2, AAAI'05. AAAI Press, Menlo Park, pp 746–751

Davis DA, Chawla NV, Christakis NA, Barabási AL (2010) Time to care: a collaborative engine for practical disease prediction. Data Min Knowl Discov 20(3):388–415. doi:10.1007/s10618-009-0156-z

Gionis A, Indyk P, Motwani R (1999) Similarity search in high dimensions via hashing. In: Proceedings of the 25th international conference on very large data bases, VLDB '99. Morgan Kaufmann Publishers Inc., San Francisco, CA, pp 518–529

Gionis A, Lappas T, Terzi E (2012) Estimating entity importance via counting set covers. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '12. ACM, New York, NY, pp 687–695

Guo Y, Greiner R (2007) Optimistic active learning using mutual information. In: Proceedings of the 20th international joint conference on artifical intelligence, IJCAI'07, pp 823–829

Hoi SCH, Jin R, Zhu J, Lyu MR (2006) Batch mode active learning and its application to medical image classification. In: Proceedings of the 23rd international conference on machine learning, ICML '06. ACM, New York, NY, pp 417–424. doi:10.1145/1143844.1143897

Ipeirotis PG, Provost FJ, Sheng VS, Wang J (2014) Repeated labeling using multiple noisy labelers. Data Min Knowl Discov 28(2):402–441

Kapoor A, Horvitz E, Basu S (2007) Selective supervision: guiding supervised learning with decision-theoretic active learning. In: IJCAI, pp 877–882

Lewis DD, Gale WA (1994) A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '94. Springer-Verlag New York Inc, New York, NY, pp 3–12

Melville P, Mooney RJ (2004) Diverse ensembles for active learning. In: Proceedings of the twenty-first international conference on machine learning, ICML '04. ACM, New York, NY, pp 74–81

Muslea I, Minton S, Knoblock C (2000) Selective sampling with redundant views. In: Proceedings of the national conference on artificial intelligence

Norén GN, Hopstadius J, Bate A, Star K, Edwards IR (2010) Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov 20(3):361–387. doi:10.1007/s10618-009-0152-3

Panigrahy R (2008) An improved algorithm finding nearest neighbor using kd-trees. In: Proceedings of the 8th Latin American conference on theoretical informatics, LATIN'08. Springer-Verlag, Berlin, Heidelberg, pp 387–398

Qian B, Li H, Wang J, Wang X, Davidson I (2013a) Active learning to rank using pairwise supervision. In: SDM, pp 297–305

Qian B, Wang X, Wang J, Li H, Cao N, Zhi W, Davidson I (2013b) Fast pairwise query selection for large-scale active learning to rank. In: ICDM, pp 607–616

Rashidi P, Cook DJ (2011) Ask me better questions: active learning queries based on rule induction. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11. ACM, New York, NY, pp 904–912. doi:10.1145/2020408.2020559

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326

Roy N, Mccallum A (2001) Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of 18th international conference on machine learning. Morgan Kaufmann, San Francisco, pp 441–448

Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison

Settles B, Craven M (2008) An analysis of active learning strategies for sequence labeling tasks. In: EMNLP, pp 1070–1079

Settles B, Craven M, Ray S (2008) Multiple-instance active learning. In: Advances in neural information processing systems NIPS. MIT Press, Cambridge, pp 1289–1296

Sun J, Wang F, Hu J, Edabollahi S (2012) Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explor 14(1):16–24

Wang F, Zhang C (2006) Label propagation through linear neighborhoods. In: Proceedings of the 23rd international conference on machine learning, ICML'06. ACM, New York, NY, pp 985–992. doi:10.1145/1143844.1143968

Wang F, Sun J, Ebadollahi S (2012) Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Stat Anal Data Min 5(1):54–69

Wang X, Wang F, Wang J, Qian B, Hu J (2013) Exploring patient risk groups with incomplete knowledge. In: ICDM, pp 1223–1228

Wauthier FL, Jojic N, Jordan MI (2012) Active spectral clustering via iterative uncertainty reduction. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '12. ACM, New York, NY, pp 1339–1347

Wu J, Roy J, Stewart WF (2010) Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. Med care 48(6):S106–S113

Zhang T, Oles FJ (2000) A probability analysis on the value of unlabeled data for classification problems. In: Proceedings 17th international conference on machine learning, pp 1191–1198

Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2003) Learning with local and global consistency. In: NIPS

Zhou J, Sun J, Liu Y, Hu J, Ye J (2013) Patient risk prediction model via top-k stability selection. In: SDM, pp 55–63

Zhu X, Ghahramani Z, Lafferty JD (2003a) Semi-supervised learning using gaussian fields and harmonic functions. In: ICML, pp 912–919

Zhu X, Lafferty J, Ghahramani Z (2003b) Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining, pp 58–65

Zhuang H, Tang J, Tang W, Lou T, Chin A, Wang X (2012) Actively learning to infer social ties. Data Min Knowl Discov 25(2):270–297