Deep Similarity-Based Batch Mode Active Learning with Exploration-Exploitation

Changchang Yin*, Buyue Qian*, Shilei Cao* Xiaoyu Li* Jishang Wei† Qinghua Zheng* Ian Davidson‡
*Xi'an Jiaotong University, Xi'an, Shaanxi, China
Email: lentery@stu.xjtu.edu.cn, qianbuyue@xjtu.edu.cn, {shileicao,wemakefocus}@stu.xjtu.edu.cn, qhzheng@xjtu.edu.cn
†HP Labs, 1501 Page Mill Rd, Palo Alto, CA 94304, USA
Email: jishang.wei@hp.com
‡Department of Computer Science, University of California, Davis, CA 95616, USA
Email: indavidson@ucdavis.edu

***Abstract*—Active learning aims to reduce manual labeling efforts by proactively selecting the most informative unlabeled instances to query. In real-world scenarios, it's often more practical to query a batch of instances rather than a single one at each iteration. To achieve this we need to keep not only the informativeness of the instances but also their diversity. Many heuristic methods have been proposed to tackle batch mode active learning problems, however, they suffer from two limitations which if addressed would significantly improve the query strategy. Firstly, the similarity amongst instances is simply calculated using the feature vectors rather than being jointly learned with the classification model. This weakens the accuracy of the diversity measurement. Secondly, these methods usually *exploit* the decision boundary by querying the data points close to it. However, this can be inefficient when the labeled set is too small to reveal the true boundary. In this paper, we address both limitations by proposing a deep neural network based algorithm. In the training phase, a pairwise deep network is not only trained to perform classification, but also to project data points into another space, where the similarity can be more precisely measured. In the query selection phase, the learner selects a set of instances that are maximally uncertain and minimally redundant (exploitation), as well as are most diverse from the labeled instances (exploration). We evaluate the effectiveness of the proposed method on a variety of classification tasks: MNIST classification, opinion polarity detection and heart failure prediction. Our method outperforms the baselines with both higher classification accuracy and faster convergence rate.**

I. Introduction

Due to the limited availability of training data and expensive human labeling cost, various active learning algorithms have been proposed to select the most informative instances from the large pool of unlabeled data. Typical active learning algorithms selected a single instance to query at each time, and then ask humans for its label. This process repeats until the learner approximately achieves the target learning accuracy or the labeling budget is reached. The key to active learning is the query strategy, whose goal is to select the most useful examples which if labeled would significantly boost the learning accuracy. Variety types of active learning methods have been proposed, such as uncertainty sampling, query by committee, and expected error reduction.

In conventional active learning, the learning model is trained too frequently with little change in the training data. This is very inefficient and would cause serious overfitting if the model is deep neural network based. To address this issue, batch mode active learning (BMAL) algorithms were proposed to select a group of instances at each iteration. In many existing BMAL methods, it's possible that a group of informative but similar instances are selected at the same time. If this is the case, it would waste the labeling effort as similar instances provide the learning model essentially the same piece of information. Therefore, besides the informativeness, diversity is the key consideration in batch mode active learning. A set of heuristic algorithms have been proposed for BMAL to choose a set of informative and diverse instances. The cluster-based algorithms [1] [2] [3] [4] firstly group the unlabeled instances based on their similarity, and then select instances in different groups to reduce the redundancy in the query set. By defining a redundancy function based on the similarity in the projected kernel space, the SVM$_{active}$ algorithms [5] [6] [2] [7] [8] select a batch of informative and diverse instances at each time. Hoi [9] proposes to choose a set of instances which have the minimum Fisher information.

However, existing BMAL algorithms mainly suffers from two limitations. (i) The performances of previous BMAL algorithms heavily rely on the accuracy of the similarity measure between instances. The algorithms map each instance into a feature space, for example the kernel space in SVM$_{active}$, and then the similarity is simply calculated based on some predefined function. The similarity amongst instances is simply calculated using the feature vectors rather than jointly learned with classification models. This weakens the accuracy of the diversity measurement. (ii) Existing BMAL algorithms are only good at "exploitation". The learners always query the instances closed to the decision boundary of the current hypothesis. At the early stage, it's probable that the number of labeled instances is too small to cover the true data distribution in the feature space. Therefore, it is also important to perform "exploration", so that the learning model would search for new regions where a large pocket of instances may be misclassified. For example, Osugi [10] cites an exclusive OR problem in Figure 1. If all the labeled data are from regions 1, 2, 3, all the data in region 4 will be misclassified. In this case, the misclassified instances are away from the decision boundary

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.67

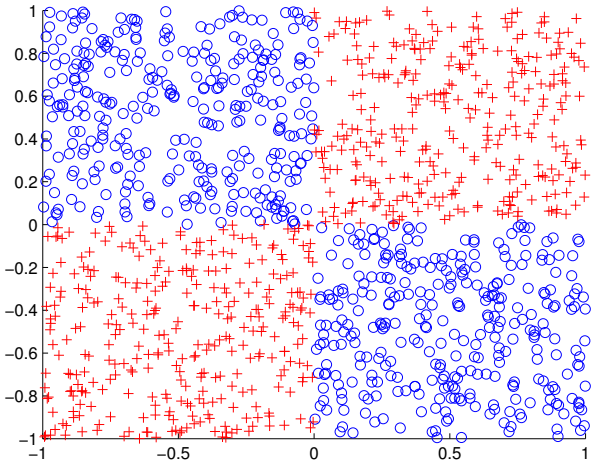

Fig. 1. An example of exclusive OR problem. We denote the upper-left, upper-right, lower-left, lower-right by 1, 2, 3, 4. Region 2 and 3 are positives and region 1 and 4 are negtives.

based on the current hypothesis, and will never be queried.

We present a novel BMAL approach that explicitly learns the similarity using deep neural network, and balances exploration and exploitation based on the learnt similarity. In our method, we adopt convolutional neural networks to perform classification. The output of the last layer (before the soft-max layer) of neural networks can be viewed as a learned feature representation of instances. Then, the similarity between a pair of instances can be calculated by the inner product of the feature vectors. Our neural network is trained with two goals in mind. (i) The first is to improve the accuracy of the classifier. (ii) The second goal is to, through the label guided feature representation learning, map the instances into another space in which the similarity can be more precisely measured.

We evaluate the proposed model on three classification tasks. The first experiment is to predict whether a patient will suffer from heart failure in the next six months. A convolutional layer and max pooling layer are used to extract features, and then a fully connected softmax layer is used to do the prediction. The second experiment is opinion polarity detection, we adopt a similar neural network as shown in [11]. Our third experiment is to classify the MNIST image with a 5-layer network. As we shall report in the experiment section, our model outperforms the baselines on all three tasks.

The main contributions of our work are as follows:

1) We propose an effective batch mode active learning scheme that takes advantages of the multi-objective capacity of deep neural network.
2) We adopt deep neural network learning methods to reform the feature representations, on which the similarity between instances can be more accurately measured.
3) We explore an Exploration-Exploitation query scheme to ensure the true diversity of instances in a batch.
4) Our method is particularly effective in the cases where the initial labeled data is in shortage, or the labeling is costly due to the requirement of specific domain knowledge. The result reported in Fig. 4 demonstrates our method in such scenario of heart failure prediction.

The contributions 2) and 3), to the best of our knowledge, are the first attempt in BMAL and have not been explored in the previous literature. The rest of the paper is organized as follows. Section II introduces related studies. Section III describes our model in detail. Section IV reports and discusses the experiment results. Section V concludes our work.

II. Related Work

In this section, we briefly review the related work. This paper is closely related to active learning and deep learning.

A. Active Learning

There are two kinds of algorithms in active learning, which are representativeness sampling and uncertainty sampling. Representativeness sampling algorithms aim to select the most representative instances according to data distribution. [1] [2] [3] [4] cluster the unlabeled instances and select the most representative instances of those clusters to query. The successes of this kind of cluster-based algorithms directly rely on the employed clustering algorithm. Huang [12] proposed a QUIRE approach which combines the informativeness and representativeness of an instance. Chattopadhyay [13] proposed to select the representative instances according to the unlabeled data distribution. Uncertainty sampling is more frequently adopted, which selects the most uncertain instances at each iteration. The uncertainty can be measured differently. Support vector machine active learning [14] chooses the instance closest to the classification boundary in kernel space. [15] [16] construct a committee, the disagreement of which can be viewed as the informativeness. All these algorithms mentioned above focused on selecting a single informative instance to query at each iteration. It's infeasible and inefficient to label the instances one by one. A mini-batch of labeled data are usually trained at the same time in the deep neural networks. It is also unreasonable to retrain the neural networks at every iteration with only one additional labeled instance. So we use batch mode active learning to select instances, which selects multiple unlabeled data simultaneously to query. There are also some works aimed to combine the representativeness sampling and the uncertainty sampling [13] [17].

B. Batch Mode Active Learning

In batch mode selection, it's crucial to reduce the redundancy between the selected instances of the same batch. Hoi [9] chooses a batch of instances which have the minimum Fisher information to reduce the redundancy. There are also many batch mode active learning algorithms based on SVM$_{active}$. Joshiy [5] proposed to greedily select a group of data with maximum utility and minimum redundancy which is measured by the probability estimates in multi-class image classification. Schohn [6] presented to measure the diversity with the distances from the separating hyperplane of a linear SVM. Brinker [2] proposed a similar algorithm to query a batch of instances based on the angles in the

hyperplane of feature space in SVM. Wang [7] combined representativeness and diversity. Xu [8] incorporated a density measure to SVM active learning. Xia [4] first clustered the unlabeled instances in feature space and then selects a batch of instances in different clusters to incorporate diversity. Guo [18] proposed a discriminative strategy that chooses a batch of points which have the minimum entropy and maximum log-likelihood. Besides, many attempts have been done to apply active learning to networked data. Zhu [19] combined active learning and semi-supervised learning to optimize harmonic functions based on Gaussian random fields. Shi [17] developed three criteria of maximum uncertainty, maximum impact and minimum redundancy, and presented an objective function combining all of them on networked data. Dasarathy [20] proposed to a simple and efficient algorithm that queries for the label of the vertex that bisects the shortest shortest path between any pair of oppositely labeled vertices.

C. Exploration and Exploitation

These algorithms are much helpful on the exploitation of refining the decision boundary by querying the instances near to the boundary. At the early stage of active learning, the small labeled set might not cover all the import regions of data space. Focusing on instances close to the decision boundary prevent exploration of the regions where the examples are currently misclassified [21]. As results, many active learning algorithms combined with exploration have been proposed. The choice of examples can be considered as the dilemma between the exploration and exploitation. Bondu [22] proposed a new strategy to manage the comprise. Osugi [10] introduced exploration to active learning by dynamically adjusting the probability to explore at each step. The update of the probability depends on the change that is induced with the newly labeled example on the hypothesis space. In the work of Cebron et al. [23], a new Prototype Based Active Learning algorithm (PBAC) was proposed, which used an integrated approach with a classification model that combines the potential of each data point and the classifier uncertainty in one single criterion. When the full space of classes is not known in advance, some previous works [24] [25] [26] [27] aim to discover the unknown classes. There are roughly two criteria of exploration. The first is to select the instances with the greatest representativeness [23]. The second is to select the instances farthest from the labeled instances [10] [28].

D. Deep learning

Most active learning algorithms combined with exploration are single mode. We propose new algorithms combining BMAL with exploration-exploitation, in which the measures of redundancy and similarity are explicitly trained in the deep neural network. Recently, deep neural networks have greatly improved the performance in many tasks, including image classification, object detection, text classification and so on. In image recognition, quite a few networks have been explored to achieve a better accuracy. Very deep convolutional networks [29] have greatly improved the image-net recognition accuracy. [30] used residual networks, which are deeper and easier to train, and performed better than [29]. Besides image recognition, deep neural networks have also achieved excellent performances. [31] [32] have driven big advances in objection detection. Zhu [33] have proposed a new model to measure the patient similarity with a deep convolutional neural network. CNN [11] has also enjoyed accuracy gains in sentiment analysis and question classification. The neural networks are becoming deeper and deeper, and need more and more data to train the models. While it's hard to get a large scale of labeled data, this paper attempts to introduce active learning to deep neural networks so as to reduce the manual labeling effort.

III. METHOD

In this part, we propose a deep similarity-based batch mode active learning algorithm, to reduce the labeling effort in deep neural networks. The algorithm can be applied to different neural networks. It aims to compute the similarity among unlabeled instances more precisely and then select the most informative instances from the unlabeled data pool to query the oracles.

A. Definition and Settings

In this paper, we focus on pool-based active learning. In the pool-based active learning setting, an algorithm actively selects data points, the labels of which are unrevealed, to requests their labels. We suppose there are a pool of u unlabeled examples $x_1, x_2, ..., x_u$ and l labeled data instances (x_{u+1}, y_{u+1}), (x_{u+2}, y_{u+2}), ..., (x_{u+l}, y_{u+l}), typically $l << u$. Let $U = \{1, 2, ..., u\}$ and $L = \{u+1, u+2, ..., u+l\}$ be the unlabeled set and labeled set respectively. We define $n = l + u$ as the total number of data instances. We denote by $y_1, y_2, ..., y_n$ the labels of $x_1, x_2, ..., x_n$. We let b denote the budget of labeled instances, which means that there are at most b instances in L.

We aim to learn a classifier $h : X \to Y$, so that the minimal generalization error $Err(h)$ is satisfied with the no more than b instances labeled, where X is the instance space and Y is the label space.

$$\mathrm{Err}(h) = \frac{1}{2|L|} \sum_{(x,y)\in L} [1 - equal(h(x), y)] \tag{1}$$

$$equal(a, b) = \begin{cases} 1 & \text{if } a = b \\ -1 & \text{else} \end{cases}$$

Active learning aims to improve the performance of the classifier h using the least number of labeled instances. The problem is that given the classifier h, labeled set L and unlabeled set U, how to select a batch of $k(k << n)$ instances S from U to label so that the quality of the classification model can be improved most. The selected instances set S should be informative and diverse at the same time. After the instances in the selected set S is labeled and added into L at each iteration, the classifier h will be retrained with the renewed L.

The algorithm repeats to select a batch of instances and then retrain the learner. There is still two question in the

TABLE I
VARIABLES USED IN OUR MODEL

Variable	Description
U	The set of unlabeled data instances
L	The set of labeled data instances
S	The set of selected data instances
i	The index of a instance
x_i	The instance i
y_i	The label of instance i
f_i	The feature vector of instance i
k	The number of instances to be selected
m	The number of instances to be selected in exploitation
b	The budget of labeled data instances

basic model. The first is how to select a batch informative instances. The second is how to retrain the classifier with the renewed L. Since the classifiers in our experiments are deep neural network models, it's very easy to cause overfitting when the label set L is small and the early labeled instances are always trained at every iteration. There are some tricks on retraining the learner. The questions are illustrated in the following subsections.

Table I gives a summary of variables used in our model.

B. Expoitation

We develop two criteria to measure the informativeness, which are maximum uncertainty and minimum redundancy.

1) Maximum Uncertainty: The type of active learning strategy is commonly known as uncertainty sampling. The learner always cares about the instances it finds confusing. The most common uncertainty sampling strategy uses entropy as the measure [34]:

$$\mathrm{E}(x) = -\sum_{0<<i<<|Y|} h_i(x) log(h_i(x)) \tag{2}$$

,where $h_i(x)$ denote the probability of that x belongs to class y_i.

The uncertainty of the selected set S is measured by summing the entropy of instances in S:

$$\mathrm{E}(S) = \sum_{i \in S} \mathrm{E}(x_i) \tag{3}$$

2) Minimum Redundancy: In conventional active learning, the instances are queried one by one, which means k = 1. The redundancy criterion is ignored. The learner always selects the instance which has the maximal entropy at each iteration. In batch mode active learning, $k > 1$. The redundancy criterion should be taken into account. The instances selected in the same batch should be diverse. It's essential to adopt an appropriate method to measure the similarity between instances. The similarities are usually measured in feature space. In SVM$_{active}$, the kernel space is used as feature space. We use the output of the last layer (before the softmax layer) as feature space. The feature vector of instance i is represented as f_i. We define the similarity function based on the feature space.

$$\mathrm{Sim}(i, j) = f_i M f_j \tag{4}$$

, where M denote a similarity matrix. If M is an identity matrix, the similarity function is represented as the product of the two feature vectors. We can also learn the M in networks, which will add some extra parameters.

$R(S)$ represents the redundancy of the selected set S :

$$\mathrm{R}(S) = \sum_{i \in S} \sum_{j \in S} \mathrm{Sim}(i, j) \tag{5}$$

To this end, we define two functions $E(S)$ and $R(S)$ that repectively represent the maximum uncertainty and minimum redundancy. The linear combination of the two functions naturally defines the objective function, i.e.,

$$\mathrm{I}(S) = \mathrm{E}(S) - \frac{\alpha}{|S|}\mathrm{R}(S) \tag{6}$$

, where α denotes a parameter to balance importances of the uncertainty and redundancy.

It's hard and very expensive to find S with the maximal $\mathrm{I}(S)$. We use a greedy algorithm instead for batch mode active selection. The exploitation part in Algorithm 1 outlines the greedy algorithm. The instance with the maximal uncertainty is firstly selected. Then the next must be mostly uncertain and different from the selected set. The similarity between an instance and the selected set S:

$$\mathrm{Sim}(i, S) = \max_{j \in S}(\mathrm{Sim}(i, j)) \tag{7}$$

Then the algorithm selects the next instance having maximum score:

$$\mathrm{I}(i) = \mathrm{E}(x_i) - \alpha \mathrm{Sim}(i, S) \tag{8}$$

C. Exploration

At the beginning of active learning, we assume that the labeled set is small and does not cover all the important regions. From the exploration point of view, we want to explore the unknown regions. Our criterion is to select the instances farthest from the labeled set. We use the formula (7) to measure the similarity between any unlabeled instance and the labeled set.

$$\mathbf{S}_e = \min_{S_e} \sum_{i \in S_e} \mathrm{Sim}(i, L \cup S) + \sum_{i \in S_e} \sum_{j \in S_e} Sim(i, j) \tag{9}$$

Solving the objective function is NP-hard. We use a greedy algorithm like in exploitation. We select the instance farthest from the labeled set each time:

$$i = \min_{i} \mathrm{Sim}(i, L \cup S) \tag{10}$$

After getting the farest instance from the L and S, we add the instance into the selected set S one by one. The concrete detail is illustrated in Algorithm 1.

Algorithm 1 greedy batch-mode selection

Input: h, U, L, k, m
Output: S
Initialize:$S = \varnothing$

Calculate entropy vector $\mathbf{E}$;
Calculate similarity matrix $\mathbf{Sim}$;
Find $i \in U$ to maximize $\mathbf{E}(x_i)$;
$S \leftarrow S \cup \{i\}$;
# exploitation
for $index = 1$ *to* $m - 1$ **do**
  for $i \in U - S$ **do**
    $\mathbf{I}(i) = \mathbf{E}(x_i) - \beta \mathbf{Sim}(i, S)$
  end for
  Find $i \in U - S$ to maximize $\mathbf{I}(i)$;
  $S \leftarrow S \cup \{i\}$;
end for
# exploration
for $index = m$ *to* $k - 1$ **do**
  for $i \in U - S$ **do**
    $\mathbf{I}(i) = -\mathbf{Sim}(i, L \cup S)$
  end for
  Find $i \in U - S$ to maximize $\mathbf{I}(i)$;
  $S \leftarrow S \cup \{i\}$;
end for

D. Combination

We will combine exploitation and exploration in batch mode active learning. We can execute exploitation and exploration respectively at each iteration as in Algorithm 1. At the beginning of active learning, it's probable that the number of labeled instances is too small to cover the true data distribution in the feature space, so the exploration should be more important. As the labeled set becomes larger and larger, the exploitation becomes the prime target. At each iteration, we will first select a set of m instances according to the exploitation criterion. Then we will select $k - m$ instances farthest from all the labeled set and the m selected instances according to the exploration criterion. We denote m as the number of instances to be selected in exploitation, which controls the importance of exploitation. At the early stage, m is relatively small. As the instances are labeled and new regions are explored, m begins to increase and the exploitation becomes more and more important. Then m can increase according to different strategies. We linearly increase m, until m reaches a predefined value m_{upper}. The experiments prove that a dynamical m works better than a static one.

E. Pairwise Neural Network

The existing algorithms usually measure the similarity among instances on the feature space. The feature space is aimed to classify the instances more precisely, but not to measure the similarity. It is difficult for the algorithms to fit the two goals at the same time. We propose a new pairwise network, the feature space of which is used to do the prediction as well as to calculate the similarity. We use deep neural networks, which have very strong learning ability and are able to fit two objective functions at the same time, to learn the classification tasks. Our model will be explicitly trained to fit both the two goals. The first goal is described in formula (1), which is to improve the accuracy of the classifier. The second goal is to more precisely map the instances into another feature space, the objective function of which can be represented as:

$$\mathrm{Err}_2(\mathbf{h}) = -\frac{\beta}{|S|^2} \sum_{i \in S} \sum_{j \in S} \mathrm{Sim}(i, j) equal(y_i, y_j) \qquad (11)$$

, where the definition of $equal(\cdot, \cdot)$ has been illustrated before. We use the parameter β to balance the two objective functions.

The network needs to be slightly modified, but will not add any extra parameter. For example, Figure 2(a) and 2(b) express the network architectures which are used to predict the onset risk of heart failure. The initial architecture is shown in Figure 2(a), which are only able to be trained to perform prediction. Our pairwise model, consisting of two initial networks which share the same variables, can learn the similarity between pairwise instances as in Figure 2(b).

In training phase, we use pairwise neural networks to learn the similarity between instances. In selection phase, the learned similarity is used to select unlabeled instances. In the test phase, the initial network is used to inference. Each of

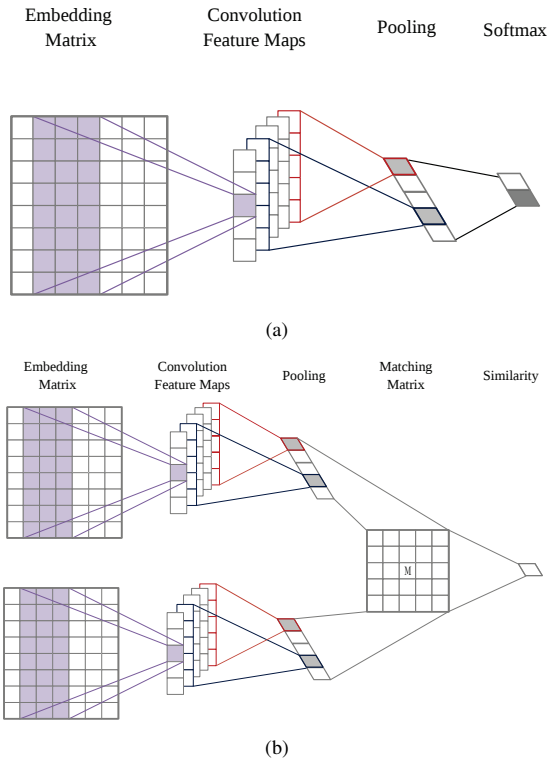

Fig. 2. The framework of heart failure prediction. (a) The original framework, which is used to predict the risk of heart failure. (b) The pairwise framework, which is aimed to learn the similarity between pairwise patients. M is used to calculate the similarity, $Sim(i, j) = f_i M f_j$. In our settings, M is an identity matrix and $Sim(i, j)$ becomes inner product.

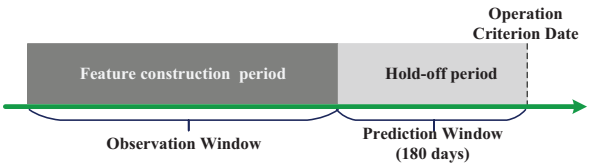

Fig. 3. Experimental setting of early prediction of the heart failure onset risk.

the networks can be trained to fit the classifier function. The pairwise network can be trained to fit the similarity function.

IV. Experiments and Evaluation

In this part, we will evaluate our approach for batch mode active learning. We compare our method to the following schemes:

- Passive random sampling, which randomly selects a batch of instances from the unlabeled pool to train the learner.
- Entropy-max sampling, which selects the top k instances with the biggest entropy as in formula (3).
- Kernel farthest first, which selects instance which is farthest from the labeled set and the selected set in the kernel space [35] [21]. We use the feature space replace the kernel space of SVM. This method is the same as the exploration part of our method in Algorithm 1.
- Batch mode active learning, which selects a batch of instances with maximum uncertainty and maximum impact [17]. The method is the same as the exploitation part of our method in Algorithm 1.

For simplicity, we use Random, EM, KFF, BMAL, Exploration to denote the above baseline methods and our method respectively.

A. Heart Failure Prediction

We conduct an experiment on a real clinical EHR(Electronic Health Records) data warehouse containing the records of 218,680 patients over four years. Each patient has a series of medical records with temporal information, which consist of demographics (i.e., age, gender, and weight), medications, procedures, lab results, diagnosis and other clinical related indicators. We will do a significant research on early prediction of heart failure, which is frequently occurred disease and extensively analyzed in healthcare applications. We carry out a case-control study on heart failure prediction. The case-control is a type of epidemiological observational study. It compares two group of subjects who have the different disease but are similar otherwise, so as to find the factors contributing to the difference more precisely. The patients confirmed with heart failure are the cases. A group of matched control patients is then collected, who have the similar demographics and characteristics.

In order to predict whether a patient will suffer from heart failure at some future time, an operation criterion date is needed for him or her. For case patients, the heart failure confirmation date is the operation criterion date. For control patients, the last day in our database is the operation criterion date. We then split each patient's EHR data into observation window and prediction window. The prediction window contains the medical records occurring at the last 180 days tracing back from the operation criterion date. The records before the prediction window, which belong to the observation window, are used for analysis.

1) Patient Selection: We construct dataset with medical events collected from patients in the EHR. We develop the criteria that the number of records of each patient in the observation window must be more than 50, so as to ensure that there are enough events of clinical history to extract medical feature to predict the diagnosis. Our domain experts help us select 4626 patients from all the satisfied, including 2323 cases and 2323 controls. Since the temporal information is so important, the medical records are reported according to the occurring date. The EHR data for every patient is then represented as a variable-length sequence. For convenience, we pad patients' records to the same size. The padding records are medically meaningless.

2) Medical Concept Embedding: It's difficult to represent the medical events of patients effectively without loss of information. Cheng [36] used a temporal matrix with time on one dimension and event on the other dimension to represent patients' EHR data. But the matrix usually is sparse and high dimensional. So we use a representation of temporal medical records with medical concept embedding similar to Zhu [33] and Zhan [37]. We train a word embedding model with the EHR data containing roughly 16.9 million medical event records of 218,680 patients. After removing the words with the frequency less than 5, there are 8627 unique medical events remained. All the medical events are converted to 50-dimensional vectors by the embedding model. The padding event's vectors are replaced with zero. The medical sequence for every patient is then converted to a fixed-size EHR matrix with each record replaced by a corresponding embedding vector.

3) Prediction Model: We implement a four-layer convolutional neural network similar to [36] to perform prediction. The first layer consists of those EHR matrices. The second layer is a one-side convolution layer extracting local features from the first layer. This layer aims to find the useful medical record patterns which are relative to heart failure. The third layer is a max pooling layer that captures the most important feature with the highest value for each feature map. The fourth layer is a fully-connected softmax prediction layer. This layer can extract global feature. The architecture of our model for heart failure onset risk prediction is presented in Figure 2(a).

In pairwise mode, the outputs of the third layer are used as features. The learner is trained alternately to fit the two objective function $\mathrm{Err}(h)$ and $\mathrm{Err}_2(h)$. The model used to learn the pairwise similarity is presented in Figure 2(b).

4) Result and Dicussion: Figure 4(a) summarizes results of 10 runs of each experiment using Random, EM, KFF, BMAL and Exploration-P(Our method). All the active learning algorithms perform better than the passive learning (Random). It demonstrates that active learning algorithms could

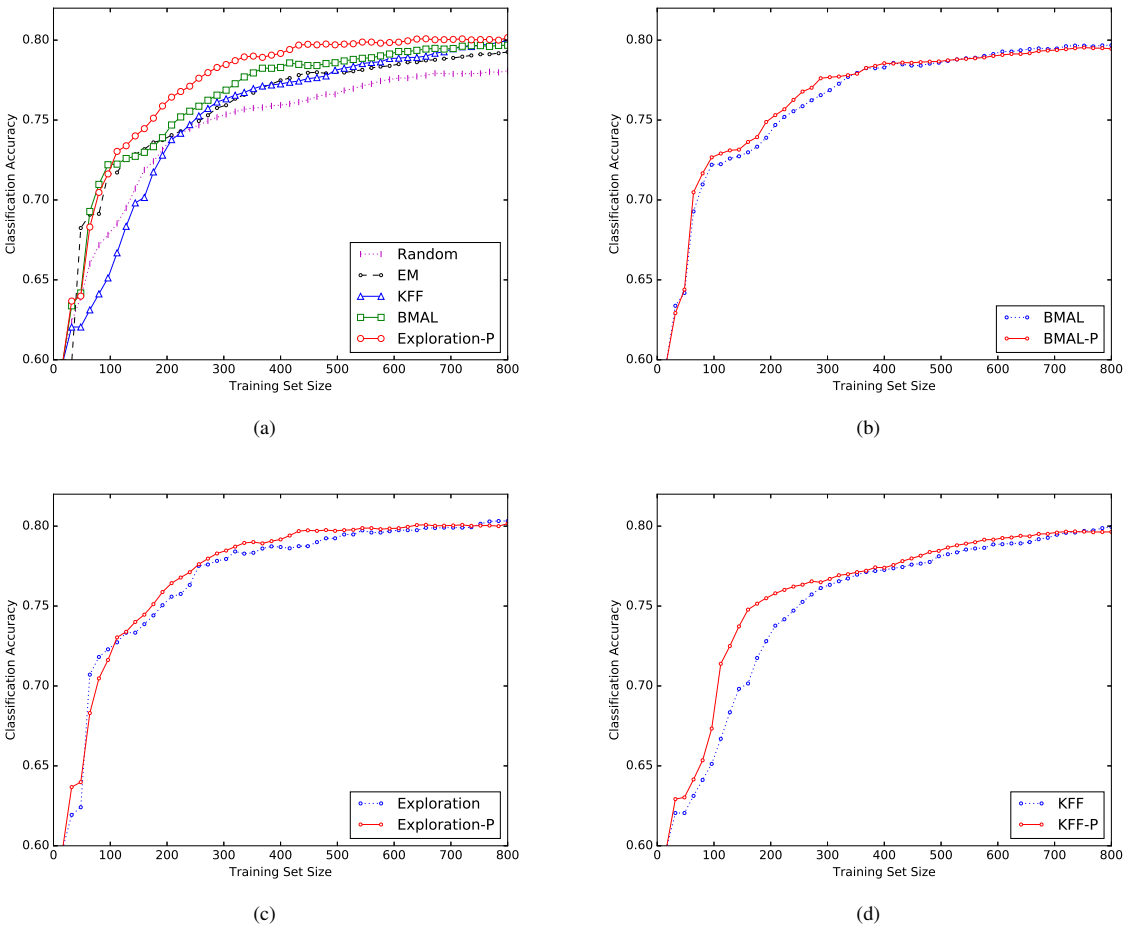

Fig. 4. Heart failure prediction results for (a) baselines and our method, (b) BMAL with and without the pairwise model, (c)Exploration with and without the pairwise model (our method), (d)KFF with and without the pairwise model.

be successfully applied to deep learning. Besides, our method with exploration and exploitation performs the best, which demonstrates the effectiveness of our method and that combining exploitation with exploration is helpful. In Figure 4(b) 4(c) 4(d), we compare the pairwise network to the original network in KFF, BMAL and Exploration. We denote "KFF", "BMAL", "Exploration" the original methods, and denote the corresponding methods with pairwise network by adding "-P" after the methods' name. We find that the all the algorithms with the pairwise network perform better than the algorithms with the original network. The precise similarity measure learned in the pairwise network cause the performance gains.

Figure 5 illustrates the optimization of hyperparameter β in our model. We find when $\beta = 0.01$, our model performs the best. The bigger of the value of β, the faster the performance improves at the beginning. It demonstrates that a precise similarity function can help find the more informative and diverse instances at the early stage. When similarity measure is relatively precise, continuing to learn the similarity function prevents to fit the classifier function. That explains why the method with the highest value of β doesn't perform very well at last.

B. Minist Classification

We use MNIST dataset [38] to validate the effectiveness of our proposed method. MNIST [38] is one of the most well-known datasets in the machine learning field. It contains 60000 images in the training set and 10000 images in the testing set. The training set and the testing set follow the same distribution. The images, with the fixed size 28*28 in gray scale, are hand-written digits ranging from 0 to 9. Examples of the hand-written digits are shown in Figure 6.

We used a simple neural network to classify the images. The network mainly consisted of three convolutions layer and two fully connected layers. There were a pooling layer and a dropout layer after each convolution layer. The output of the first fully connected layer is a 625-dimensional vector for each image, which can be viewed as the image's feature. The output of the network for each image is a 10-dimensional vector corresponding to 10 classes. We use ReLU as activation

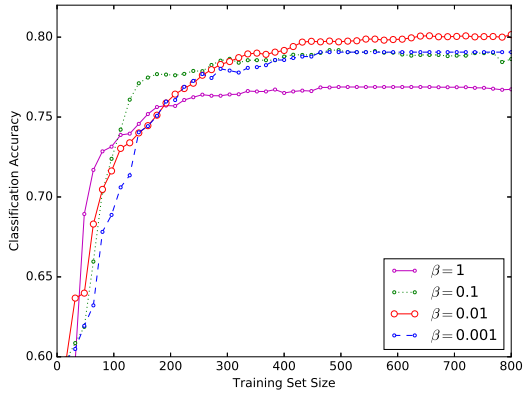

Fig. 5. Active learning results of different β for the onset risk prediction of heart failure. Results are averaged over 10 runs.

Fig. 6. Examples from the MNIST database.

function throughout the network. We apply our active learning algorithm to the model. We connect two same networks, which share all the variables, by multiplying the output feature of the first fully connected layer. The output of the pairwise network implies the similarity between the pairwise input images.

We run the network 10 times. The average results are shown in Figure 7 and 8. At the early stage, passive learning performs better than some active learning methods. We conjecture that the training set selected by active learning is very unbalanced, while the passive learning selects the training set with the same distribution as the real dataset, which is balanced. Then active learning algorithms achieve faster convergences and lead to better generalization. The batch mode algorithm outperforms the entropy-max algorithm, which demonstrates the effectiveness of redundancy criterium.

Our method combines KFF and BMAL. At the beginning, our method selects most instances in S with KFF, so our method and KFF perform almost identically. Because they are both focused on exploration, their performances are much better than others. When training set size becomes larger and larger, our method selects more and more instances with BMAL rather than KFF. Since our method focused more on exploitation, at last, it starts to surpass KFF. The results demonstrate that the exploration is more important at the early stage while the exploitation is more important at the late stage.

We add some noise to the dataset to test whether the algorithm is robust. When active learner selects the instances to query, the labels are randomly changed with probability $\eta \in \{0, 0.05, 0.10, 0.15\}$. Figure 7 shows the results of the methods under different noise rate. We found that our algorithm is robust against high rates of classification noise.

Because the networks possibly forget the data trained before, the model should be trained with all the selected data. The ratio of the newly selected data to all the labeled data is influential in the performance of the methods. We train the model with the previously labeled data and new labeled data meanwhile, and each instance has the corresponding probability of being chosen to train the model at each iteration. Figure 8 shows the results under a different ratio of the new labeled data of each mini-batch in training phase. We denote r the ratio of the number of newly selected data to batch size. When r = 1, only newly selected data are used to train the model. The performances of all the methods significantly degenerate except Random. We argue that the cause is that the model forgets the previous instances and always focuses on the new selected instances which are usually extremely unbalanced. The random sampling performs relatively well because of the balanced selected set. When r = 0, all the labeled instances have the same probability to be chosen at each mini-batch. We find that when newly selected data and previously selected data are mixed properly, the model has the best performance.

C. Opinion Polarity Detection

We also evaluate our algorithm on the opinion polarity detection subtask of the MPQA dataset [39]. The dataset consists of 10662 sentences, including 5331 positive sentences and 5331 negative sentences. We use the model in [11] to classify the opinion polarity.

Figure 9 shows the results of different active learning algorithms. We see that our proposed method still performs the best in text classification. The KFF method has a pretty good performance at the early stage, which means that the exploration is important at the beginning of the learning process. The BMAL method outperforms the KFF method at last, which implies that almost all the important regions are explored and the exploitation becomes essential to improve the classification accuracy.

V. Conclusion

In this paper, we address the problem of batch mode active learning. We propose a new deep-similarity based batch mode active learning algorithm and successfully combine active learning algorithms with deep neural networks in this paper. We use a pairwise deep neural network to train the feature representations of instances, so as to maximize the diversity of selected data in selection phase. The experimental results show that our pairwise networks achieve better similarity measures which cause better batch selection than the original single-mode networks. Based on the precise similarity measure, we propose a joint exploration-exploitation BMAL algorithm which combines BMAL and KFF. We use a parameter β to balance the importances of exploitation and exploration. By appropriately combining exploration and exploitation, the learner can make the selected batch of instances more diverse and informative. The experimental results show that our joint exploration-exploitation BMAL significantly outperforms the baselines. In future work, we plan to measure the "success"

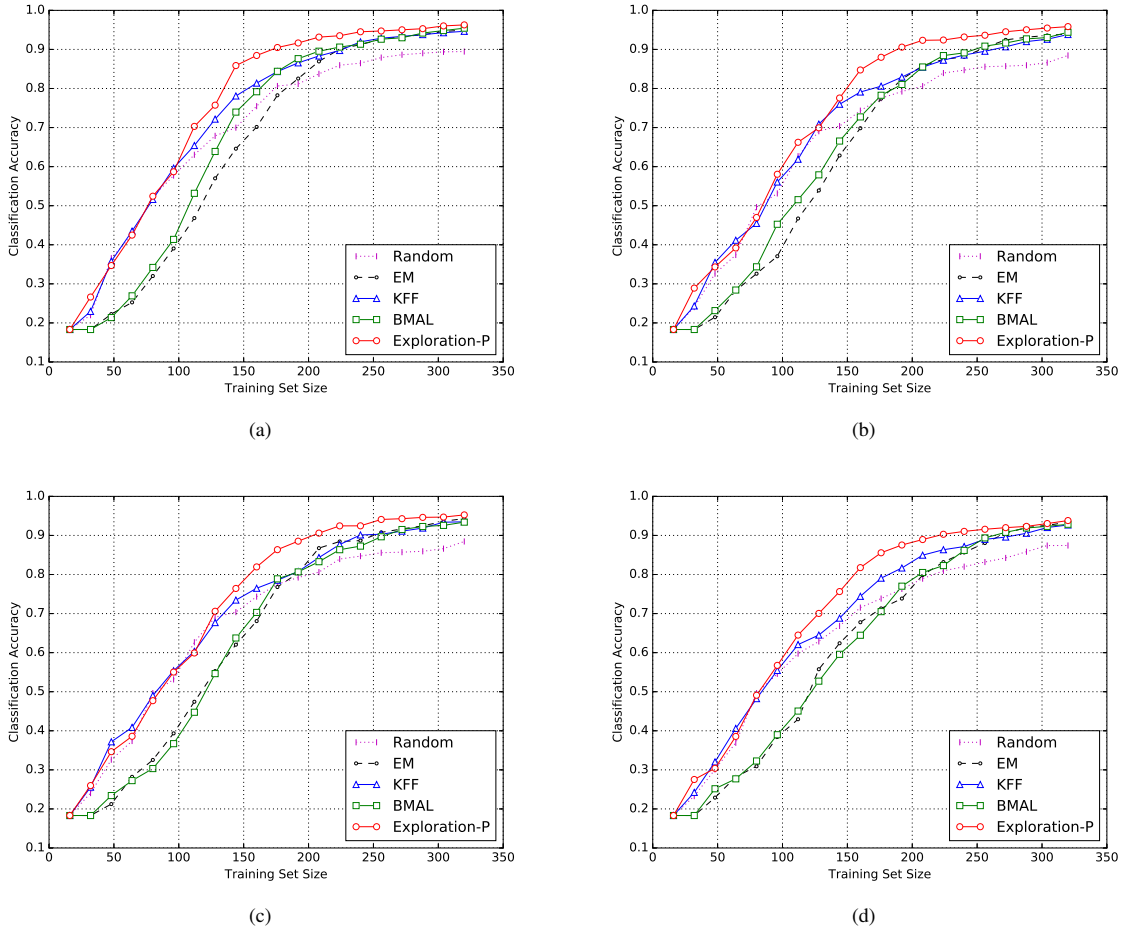

Fig. 7. MNIST classification results for (a) no noise, (b) noise rate = 0.05, (c) noise rate = 0.10, and (d) noise rate = 0.15.

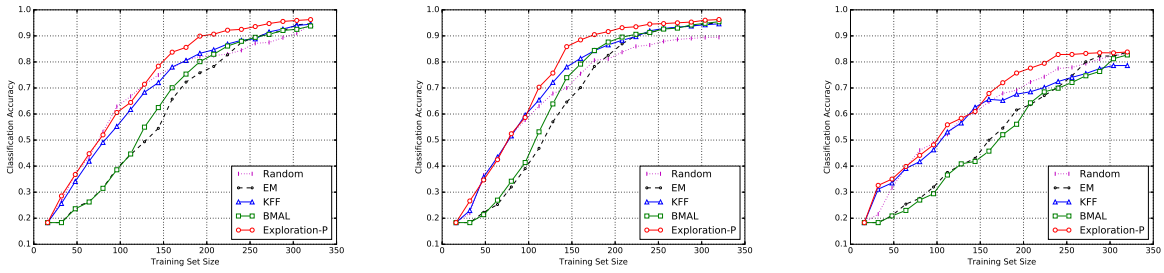

Fig. 8. MNIST classification results for (a) new data rate r = 0, (b) new data rate r = 0.5, (c) new data rate r = 1.0. Let r denote the ratio of the number of new selected data to batch size in training phase.

of the exploration at each step, which means how much the exploration contributes to the accuracy gain of the classifier. In addition, we plan to study how to dynamically adjust the importance of exploration according to the "success" of the exploration.

ACKNOWLEDGMENT

This work is sponsored by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with grant number 2016YFB1000903; Project of China Knowledge Centre for Engineering Science and Technology;

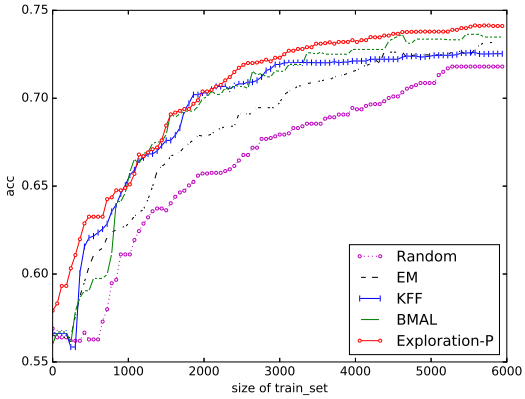

Fig. 9. Active learning results of the opinion polarity dataset.

National Natural Science Foundation of China Innovation Research Team No. 61721002; Ministry of Education Innovation Research Team No. IRT_17R86; Project of China Knowledge Centre for Engineering Science and Technology; "Multi-model based Patient similarity learning for medical data modeling and learning" under National Natural Science Foundation of China General Program.

REFERENCES

[1] H. T. Nguyen and A. W. M. Smeulders, "Active learning using pre-clustering," p. 79, 2004.

[2] K. Brinker, "Incorporating diversity in active learning with support vector machines," pp. 59–66, 2003.

[3] "Advances in knowledge discovery and data mining, 8th pacific-asia conference, PAKDD 2004, sydney, australia, may 26-28, 2004, proceedings," vol. 3056, 2004.

[4] X. Xia, P. Protopapas, and F. Doshivelez, "Cost-sensitive batch mode active learning: Designing astronomical observation by optimizing telescope time and telescope choice," pp. 477–485, 2016.

[5] A. J. Joshiy, F. Porikli, and N. Papanikolopoulos, "Multi-class batch-mode active learning for image classification," pp. 1873–1878, 2010.

[6] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," pp. 839–846, 2000.

[7] Z. Wang, B. Du, L. Zhang, and L. Zhang, "A batch-mode active learning framework by querying discriminative and representative samples for hyperspectral image classification," *Neurocomputing*, vol. 179, pp. 88–100, 2016.

[8] "Advances in information retrieval, 29th european conference on IR research, ECIR 2007, rome, italy, april 2-5, 2007, proceedings," ser. Lecture Notes in Computer Science, G. Amati, C. Carpineto, and G. Romano, Eds., vol. 4425. Springer, 2007.

[9] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," pp. 417–424, 2006.

[10] T. Osugi, D. Kim, and S. Scott, "Balancing exploration and exploitation: a new algorithm for active machine learning," pp. 330–337, 2005.

[11] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1746–1751.

[12] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," pp. 892–900, 2010.

[13] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Batch mode active sampling based on marginal probability distribution matching," in *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, 2012, pp. 741–749.

[14] S. Tong and E. Y. Chang, "Support vector machine active learning for image retrieval," pp. 107–118, 2001.

[15] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, no. 2-3, pp. 133–168, 1997.

[16] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," pp. 287–294, 1992.

[17] L. Shi, Y. Zhao, and J. Tang, "Batch mode active learning for networked data," *ACM TIST*, vol. 3, no. 2, pp. 33:1–33:25, 2012.

[18] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," pp. 593–600, 2007.

[19] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," 2003.

[20] G. Dasarathy, R. D. Nowak, and X. Zhu, "S2: an efficient graph based active learning algorithm with application to nonparametric classification," in *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 2015, pp. 503–522.

[21] Y. Baram, R. Elyaniv, and K. Luz, "Online choice of active learning algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 255–291, 2004.

[22] A. Bondu, V. Lemaire, and M. Boulle, "Exploration vs. exploitation in active learning : A bayesian approach," pp. 1–7, 2010.

[23] N. Cebron and M. R. Berthold, "Active learning for object classification: from exploration to exploitation," *Data Mining and Knowledge Discovery*, vol. 18, no. 2, pp. 283–299, 2009.

[24] T. S. F. Haines and T. Xiang, "Active learning using dirichlet processes for rare class discovery and classification." 2011.

[25] C. C. Loy, T. Xiang, and S. Gong, "Stream-based active unusual event detection," pp. 161–175, 2010.

[26] C. C. Loy, T. M. Hospedales, T. Xiang, and S. Gong, "Stream-based joint exploration-exploitation active learning," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 1560–1567.

[27] J. W. Stokes, J. C. Platt, J. Kravis, and M. Shilman, "Aladin: Active learning of anomalies to detect intrusions," *Microsoft Research*, 2008.

[28] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2001.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[31] R. B. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1440–1448.

[32] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[33] Z. Zhu, C. Yin, B. Qian, Y. Cheng, J. Wei, and F. Wang, "Measuring patient similarities via a deep architecture with medical concept embedding," in *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, 2016, pp. 749–758.

[34] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[35] D. S. Hochbaum and D. B. Shmoys, "A best possible heuristic for the k-center problem," *Mathematics of Operations Research*, vol. 10, no. 2, pp. 180–184, 1985.

[36] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," pp. 432–440, 2016.

[37] M. Zhan, S. Cao, B. Qian, S. Chang, and J. Wei, "Low-rank sparse feature selection for patient similarity learning," in *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, 2016, pp. 1335–1340.

[38] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[39] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.