Neurocomputing 73 (2010) 2980–2988

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Active learning with extremely sparse labeled examples ☆

Shiliang Sun [a,*], David R. Hardoon [b]

[a] Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, China
[b] Data Mining Department, Institute for Infocomm Research (I^2R), A*STAR, 1 Fusionopolis Way, #20-10 Connexis, Singapore

ARTICLE INFO

Article history:
Received 15 June 2009
Received in revised form
12 October 2009
Accepted 26 July 2010
Communicated by C. Fyfe
Available online 26 August 2010

Keywords:
Active learning
Multi-view learning
Canonical correlation analysis
Text classification
Image classification

ABSTRACT

In the setting of active learning there exists a general assumption that labeled examples are available for training a classifier, which in turn is used to examine unlabeled data to select the most 'informative' examples for manual labeling. However, in some domain applications there are a limited number of labeled examples available, such as in the most extreme cases of having a single labeled example per category. In these scenarios, the most existing active learning methodologies cannot be directly applied without initially making an assumption on label assignment. In this paper we present a method for finding high-informative examples for manual labeling based on extremely limited labeled data available during training. We propose using canonical correlation analysis to investigate the correlation between different views of the available data and demonstrate that this measure can be used as a selection criterion for the novel application of active learning using only a single labeled example from each class. We demonstrate our method with promising experimental results on text classification, advertisement removal and multi-class image classification tasks.

1. Introduction

In supervised learning algorithms we require a number of labeled examples for training a classifier. However, there exist a number of machine learning and data mining applications, where labeling examples is difficult, expensive or a time consuming process due to the requirement of large amounts of, what can be tedious, labor from experienced annotators. As a consequence, there generally exists a much larger set of unlabeled examples than labeled ones. For example, in web-page classification or content-based image retrieval we can easily build a large database of documents or images, but labeling a considerable portion of the database is almost infeasible when considering time and cost.

Active learning is the scenario, where the learning algorithm can actively query the user for labels. Due to this interactive setting, the number of examples needed to learn a concept can often be much lower than the number of examples required in a normal supervised learning setting. In other words, we can view the aim of active learning as reducing the number of examples needed to be labeled or generated by investigating the values of different examples. Active learning can be roughly divided into two categories: selective sampling [18] and interventional experimentation [24]. In selective sampling an active learner selects the most informative candidates, following some criterion, from a large pool of unlabeled data for human labeling. Interventional experimentation can force interesting variables to be set to certain values given an experiment, for example, feeding a rat with food which is not normally eaten by rats to observe the values of interesting variables [24]. This paper focuses on the selective sampling setting for active learning.

In order for active learning methods to operate, a number of labeled examples are usually needed for a moderate classifier to be trained, which is then used on the unlabeled data to select the most informative examples for labeling. Nonetheless, in some applications the number of attainable labeled examples is extremely limited, which in turn hampers our ability to construct an efficient classifier. In particular, there may be only one labeled example from each category. For example, consider an online product-recommendation system for anonymous internet visitors (or online web-page recommendation). When a visitor browses an interesting product, the system can try to provide them products of similar content. This can be simplified to a binary classification problem. The product, the visitor is browsing, is the positive example and the system can easily find a negative example by providing the visitor with products of various types and asking for a single feedback on the product the visitor is not interested in. In this case, the system has obtained one labeled example from the positive and negative classes, respectively. Similarly in content-based image retrieval the query image will constitute the positive example and the retrieval systems can offer the user images of various types to obtain a negative example. In these scenarios, if the initial classifier cannot be

☆This paper extends the abstract with a similar title presented at the NIPS 2008 Workshop on Learning from Multiple Sources.

* Corresponding author. Tel.: +86 21 54345186; fax: +86 21 54345119.
E-mail addresses: slsun@cs.ecnu.edu.cn, shiliangsun@gmail.com (S. Sun).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.07.007

learned from the limited labeled data available the corresponding active learning methods would not work well.

This paper proposes active learning with extremely sparse labeled examples (ALESLE), which works under a multi-view setting. Specifically, for each considered learning problem, we assume there is only one example labeled from each class. By exploiting the correlation between features of the different views, additional high-informative examples can be selected for labeling and then incorporated in subsequent active learning algorithms. Experiments in different domains show the effectiveness of the proposed ALESLE method.

It should be noted that although we have used the high-informative examples found by our proposed approach to a subsequent active learning context, they can also be used to other learning scenarios. The remainder of the paper is organized as follows. In Section 2 we give a brief review on related work and our motivation, furthermore we clarify the contribution of the proposed method, whereas in Section 3 we describe ALESLE in detail. Section 4 applies the proposed method to multiple real word data sets and reports the experimental results. Finally, in Section 5 we bring forward our conclusions.

2. Related work and motivation

The earliest works on active learning can be attributed to the contribution of Angluin [1] who had considered the problem of using queries and answers to learn an unknown concept in a formal framework. Several types of queries were described and studied including membership, equivalence, subset, superset, disjointness and exhaustiveness. Angluin also described some efficient algorithms and lower bounds on the number of queries in several domains.

Until now, there have been a number of selective sampling methods proposed in the literature, although some of them can also be applied to interventional experimentation. The proposed active learning methods can be subdivided into two major paradigms: uncertainty sampling and committee-based sampling [18]. Uncertainty sampling selects for labeling the unlabeled data on which the learned single classifier is the least confident. Representative methods of uncertainty sampling include studies in [5,11,16,25]. Cohn et al. [5] reviewed the use of optimal data selection techniques with feedforward neural networks, and further extended the principles to active learning with two statistical learning architectures: mixtures of Gaussians and locally weighted regression. Hoi et al. [11] introduced a framework for batch mode active learning where Fisher information matrix is adopted to simultaneously select some informative examples. Lewis and Gale [16] applied a probabilistic classifier to a text classification task, and selected the examples for manual labeling based on their posterior probabilities. Tong and Koller [25] presented an algorithm for performing active learning with support vector machines. Theoretically, this algorithm is well motivated in terms of the concept of version space[1] [17].

Committee-based sampling measures the degree to which a committee of classifiers disagrees. It selects for labeling the unlabeled examples whose classification is the most uncertain amongst the committee member classifiers. Representative methods include work in [6,21,26]. Freund et al. [6] analyzed the query by committee algorithm proposed in [21] in the Bayesian learning framework and showed that for some natural learning problems, prediction errors decrease exponentially fast with the number of queries. Zhou and Goldman [26] presented a democratic priority sampling method which integrates the confidence of each individual classifier (committee member) in priority estimates. While query by committee [21] used a vote entropy to determine the unlabeled examples for active labeling, the democratic priority sampling employs a confidence-weighted vote entropy to implement this task.

Unlike the described active learning methods, co-testing is a multi-view active learning method, which repeatedly trains a classifier from each view and queries contention points from unlabeled examples where the two classifiers have different predictions for labeling [18]. Co-testing has been shown to have intrinsic superiority over the single-view uncertainty and committee-based sampling approaches [18]. Recently, multi-view active learning has also been applied to information extraction [13].

Current active learning methods usually require a number of labeled training examples before they can learn a classifier to further examine the unlabeled data. When encountering applications where only a limited number of labeled examples are available, the most current active learning methods cannot be applied. It is important to note that while relevance feedback process, in content-based image retrieval, can be generated from a single query (labeled positive example), it is not a genuine active learning method as it does not provide the most informative examples for users to annotate. Furthermore, relevance feedback has limitations in sampling examples to be labeled as it is susceptible to select redundant and uninformative examples [16]. Our proposed multi-view active learning method ALESLE can work in the situation where extremely limited data are available for training, which to the best of our knowledge is the first systematical contribution on active learning from sparse labeled training data.

Our present study is largely motivated by [3,22,27]. Blum and Mitchell [3] had noted that there are many real world domains with natural multiple view data. For example, in web-page classification, a web page can be described by the text appearing on the document itself and by the anchor text attached to the hyperlinks pointing to this page. In television broadcast understanding, broadcast segments can be simultaneously described by their video and audio signals. Further examples include web-image retrieval, where an image can be described by its visual information or by the surrounding text. Theoretically, co-training assumes that features from each view are sufficient to train a good classifier, and that the two views are conditionally independent given the class [3]. In a recent study, Balcan et al. [2] relaxed the conditional independence assumption with a much weaker expansion condition given appropriately strong PAC-learning algorithms on each view.

Previously, Sun et al. [22] introduced the transduction of labeled examples (TLE) method that works with one available labeled example from each category. TLE was applied in the context of semi-supervised learning, a field which concerns on how to combine labeled and unlabeled examples to train a good classifier where manual labeling is not involved [4]. The TLE method seeks the mode of the example distribution corresponding to each labeled example by mean shift [7]. Mean shift makes use of the shift of sample means to estimate the gradient of a distribution, and thus can be used to determine the mode of the distribution. Each example at the mode is then regarded as an extra labeled example for classification. This method cannot work well in high-dimensional spaces with comparatively limited training data as, in this scenario, the modes found would be unreliable.

The one labeled example and two views (OLTV) method proposed by Zhou et al. [27] is a semi-supervised learning method, which assumes that there only exists a single positive labeled example. In order to make the use of subsequent semi-supervised learning methods, it uses kernel canonical component

[1] Given a set of labeled training data, the set of hypotheses which are consistent with the data is called the version space.

analysis [8] to increase the number of labeled data by selecting unlabeled examples with the highest and lowest similarity scores (automatically regarded as positive and negative examples) to the labeled example. Since OLTV only uses one labeled example from the positive class, the examples labeled from the unlabeled data may be unreliable. A further and important limitation of OLTV is that it cannot be directly applied to multi-class discrimination. In addition, the limitation of only using one positive labeled example makes it impossible to apply active learning algorithms as it is hard to determine informative unlabeled examples.

The proposed ALESLE method works in the active learning context and unlike TLE and OLTV, which have been proposed for semi-supervised learning, ALESLE can be naturally applied in high-dimensional spaces and in multi-class discrimination scenarios.

3. ALESLE

3.1. Problem setting

We begin by giving the nomenclature used throughout the paper. Let $\mathcal{V}=\mathcal{V}^1\times\mathcal{V}^2$ be the instance space in the two-view background where $\mathcal{V}^1$ and $\mathcal{V}^2$ correspond to two different views (features sets) of an example. Let $(\langle\mathbf{x},\mathbf{y}\rangle,z)$ denote a labeled example where vector $\mathbf{x}$ and $\mathbf{y}$ are features, respectively, arising from space $\mathcal{V}^1$ and $\mathcal{V}^2$, and scalar $z\in\mathbb{N}$ is a class label. In this paper, we first limit ourselves to the binary classification with $z\in\{1,-1\}$ (for binary classification we use a different representation for the range of z), and then extend our method to the multi-class case.

Co-training assumes each view is sufficient for correct classification [3]. Formally, for examples with non-zero probability there exist functions $f_{\mathcal{V}}$, $f_{\mathcal{V}^1}$ and $f_{\mathcal{V}^2}$ defined on the corresponding spaces indicated by subscripts, such that

$$f_{\mathcal{V}}(\langle\mathbf{x},\mathbf{y}\rangle)=f_{\mathcal{V}^1}(\mathbf{x})=f_{\mathcal{V}^2}(\mathbf{y})=z.$$

We assume that there is only one labeled example from each class, therefore the two labeled examples are, respectively, defined as $(\langle\mathbf{x}_1,\mathbf{y}_1\rangle,1)$ and $(\langle\mathbf{x}_2,\mathbf{y}_2\rangle,-1)$. Finally, the remaining unlabeled data set is $\mathcal{U}=\{(\langle\mathbf{x}_i,\mathbf{y}_i\rangle,z_i)\}$ for $i=3,\ldots,l$ where the true label z_i is unknown. Our task is to induce a classifier for classifying new data based on the two labeled examples and those selected and manually labeled from $\mathcal{U}$.

3.2. Methodology

Motivated by the assumption that the two views are sufficient for correct classification [3] we further assume that each view should have some close relationship with the semantic characteristic of underlying patterns we aim to learn. Therefore, the two views must be strongly correlated in some way [10,27]. As correlation analysis relies on the coordinate system adopted to describe variables, we attempt to find projections, by linear transformations, of the two views which can unfold the latent correlation. The correlated projections can help determine the labels of the unlabeled data. In particular, canonical correlation analysis (CCA) is used to identify the projections [12] and a similarity measure in the space of correlated projections is adopted to judge the similarity of unlabeled examples to the original labeled data.

We briefly review CCA for completeness of the presented study. CCA finds basis vectors for the two feature sets, respectively, from each of the two views, such that the projections of these feature sets to the basis vectors are maximally correlated [9,10,12]. For the current problem, CCA tries to find two basis vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ for feature matrices

$$\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_l)$$

and

$$\mathbf{Y}=(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_l)$$

in order to maximize the correlation coefficient between projections $\mathbf{w}_x^\top\mathbf{X}$ and $\mathbf{w}_y^\top\mathbf{Y}$. Let $\mathbf{C}_{xy}$ denotes the between-sets covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ while $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ denote the within-sets covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The objective function for CCA is

$$\max_{\mathbf{w}_x,\mathbf{w}_y}\ \mathbf{w}_x^\top\mathbf{C}_{xy}\mathbf{w}_y \quad \text{s.t.}\ \begin{cases}\mathbf{w}_x^\top\mathbf{C}_{xx}\mathbf{w}_x=1\\ \mathbf{w}_y^\top\mathbf{C}_{yy}\mathbf{w}_y=1.\end{cases}$$

The corresponding Lagrangian is

$$L(\lambda_x,\lambda_y,\mathbf{w}_x,\mathbf{w}_y)=\mathbf{w}_x^\top\mathbf{C}_{xy}\mathbf{w}_y-\frac{\lambda_x}{2}(\mathbf{w}_x^\top\mathbf{C}_{xx}\mathbf{w}_x-1)-\frac{\lambda_y}{2}(\mathbf{w}_y^\top\mathbf{C}_{yy}\mathbf{w}_y-1).$$

Taking derivatives to $\mathbf{w}_x$ and $\mathbf{w}_y$, and letting them equal to zero, we obtain

$$\partial L/\partial\mathbf{w}_x=\mathbf{C}_{xy}\mathbf{w}_y-\lambda_x\mathbf{C}_{xx}\mathbf{w}_x=0 \tag{1}$$

$$\partial L/\partial\mathbf{w}_y=\mathbf{C}_{yx}\mathbf{w}_x-\lambda_y\mathbf{C}_{yy}\mathbf{w}_y=0. \tag{2}$$

Subtracting $\mathbf{w}_y^\top\times(2)$ from $\mathbf{w}_x^\top\times(1)$ results in

$$\lambda_y\mathbf{w}_y^\top\mathbf{C}_{yy}\mathbf{w}_y-\lambda_x\mathbf{w}_x^\top\mathbf{C}_{xx}\mathbf{w}_x=\lambda_y-\lambda_x=0.$$

Let $\lambda_y=\lambda_x=\lambda$. For invertible $\mathbf{C}_{yy}$, from (2) we get

$$\mathbf{w}_y=\frac{1}{\lambda}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x. \tag{3}$$

Substituting (3) into (1) results in the following generalized eigenvalue problem:

$$\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x=\lambda^2\mathbf{C}_{xx}\mathbf{w}_x. \tag{4}$$

Then $\mathbf{w}_x$ can be solved, and the corresponding $\mathbf{w}_y$ would be obtained from (3). Now the objective function would be

$$\mathbf{w}_x^\top\mathbf{C}_{xy}\mathbf{w}_y=\frac{1}{\lambda}\mathbf{w}_x^\top\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{w}_x=\lambda\mathbf{w}_x^\top\mathbf{C}_{xx}\mathbf{w}_x=\lambda. \tag{5}$$

From (5), it is clear that eigenvectors corresponding to large eigenvalues in (4) should be chosen. Finally, there is often a normalization procedure to make $\mathbf{w}_x$ and $\mathbf{w}_y$ be unit vectors, as we only concern projection directions.

From (4), more than one $\mathbf{w}_x$ with the corresponding λ can be identified and consequently multiple $\mathbf{w}_y$ would be obtained. The degree of correlation between projections is reflected by λ. For real applications, it is often the case that there are a lot of basis vector pairs $(\mathbf{w}_x,\mathbf{w}_y)$ available to reflect different correlations. Assume CCA identifies m pairs of correlated projections and the corresponding correlation coefficients are λ_j for $j=1,\ldots,m$. An example $\langle\mathbf{x},\mathbf{y}\rangle$ will be transformed to m projection pairs denoted as $\langle P_j(\mathbf{x}),P_j(\mathbf{y})\rangle$ for $j=1,\ldots,m$. Given an labeled example $\langle\mathbf{x}_k,\mathbf{y}_k\rangle$ for $k=1,2$ and an original unlabeled example $\langle\mathbf{x}_i,\mathbf{y}_i\rangle$ for $i=3,\ldots,l$, the similarity in the jth projection between these two examples can be calculated as

$$S_{j,i}^k:=\exp(-(P_j(\mathbf{x}_i)-P_j(\mathbf{x}_k))^2)+\exp(-(P_j(\mathbf{y}_i)-P_j(\mathbf{y}_k))^2).$$

The total similarity between these two examples is given by

$$\rho_i^k=\sum_{j=1}^{m}\lambda_j S_{j,i}^k. \tag{6}$$

Since the effectiveness of this similarity measure for determining similarities and its superiority over the k-nearest neighbor rule

have been shown in OLTV [27] and our previous study [23], we also employ this measure in ALESLE.

Following our assumption that there is only one labeled example from each category, the ALESLE method uses the above similarity rule to calculate d values and select unlabeled examples with small d values, where d has the following definition for an unlabeled example with index i:

$$d_i = \|\rho_i^1 - \rho_i^2\|. \tag{7}$$

The intuition is that examples with small d values tend to be equidistant from the two labeled examples, and lie somewhere at the boundary between these two classes. Therefore, labeling these examples would be very helpful. As a consequence, the number of labeled training data would increase allowing for general active learning methods to be applied. Since the distance metric in (7) uses semantic information in the space of correlated projections, the induced examples may be more useful than those obtained by the general manner of random selection from $\mathcal{U}$. In addition, to retain as much semantic information as possible the parameter m in (6) is taken to be the number of eigenvalues λ^2 in (4) whose values are larger than 0.01. This corresponds to keep all correlation coefficients with λ greater than 0.1. We give the pseudocode of the ALESLE method in Table 1. Considering the total data distribution and existence of noise, it does not select the n examples with least d_i values directly from $\mathcal{U}$.

Table 1
ALESLE pseudocode.

Input:
$\mathcal{L} = \{(\langle \mathbf{x}_1, \mathbf{y}_1 \rangle, 1), (\langle \mathbf{x}_2, \mathbf{y}_2 \rangle, -1)\}$; $\mathcal{U} = \{(\langle \mathbf{x}_i, \mathbf{y}_i \rangle, z_i)\}$ $(i = 3, \ldots, l)$.
m: number of pairs of correlated projections retained.
n: number of actively selected examples for labeling.
Algorithm:
1 Obtain $\mathbf{w}_x$, $\mathbf{w}_y$ and λ according to (4) and (3); keep m pairs of basis vectors with largest λ.
2 **For** $i = 1, \ldots, l$ Project $\langle \mathbf{x}_i, \mathbf{y}_i \rangle$ to the m pairs of basis vectors.
3 **For** $i = 3, \ldots, l$ Compute d_i according to (7).
4 Choose a subset $\mathcal{A}$ from $\mathcal{U}$ with least d_i values.
5 Label n randomly selected examples from $\mathcal{A}$. The n labeled examples constitute set $\mathcal{A}_L$.
6 $\mathcal{L} = \mathcal{L} + \mathcal{A}_L$, $\mathcal{U} = \mathcal{U} - \mathcal{A}_L$.
Output: $\mathcal{L}$, $\mathcal{U}$.

The extension of ALESLE to multi-class discrimination is straightforward. Suppose M classes are considered and M examples are provided as original labeled ones each of which belongs to one of the M classes. According to (6), for an unlabeled example $\langle \mathbf{x}_i, \mathbf{y}_i \rangle$ we can obtain M similarities ρ_i^k for $k = 1, \ldots, M$. Then, we calculate d_i as follows:

$$d_i = \sum_{k=1}^{M} \|\rho_i^k - \overline{\rho}_i\|, \tag{8}$$

where $\overline{\rho}_i = \frac{1}{M}\sum_{k=1}^{M} \rho_i^k$. When $M=2$ (for binary classification), (8) degenerates to (7). In other words, the multi-class extension includes the binary classification as a special case.

ALESLE allows for more informative labeled examples to be obtained, and thus the subsequent execution of general active learning methods. Since the proposed ALESLE method works under a two-view setting, naive co-testing is used to implement subsequent active learning tasks. Naive co-testing is a state-of-the-art multi-view active learning method [18] which, based on the labeled informative examples provided by ALESLE, trains two classifiers, respectively, from each view and then randomly selects one of the contention points of the two classifiers for further manual labeling. With the progress of this active learning, more and more informative examples are labeled which in turn benefit the training of accurate classifiers. Although here naive co-testing is used to combine with ALESLE, in fact other active learning methods can also be employed.

4. Experiments

In this section, we describe a number of experiments, where ALESLE is used for active data labeling including text classification, advertisement removal and multi-class image classification. For the first two tasks, a 10×10-fold cross validation (CV) is performed. In each division, one positive and negative examples are randomly selected from the nine training folds to be used as the labeled training examples. The remaining data within the training folds are used as the unlabeled data set. The performance

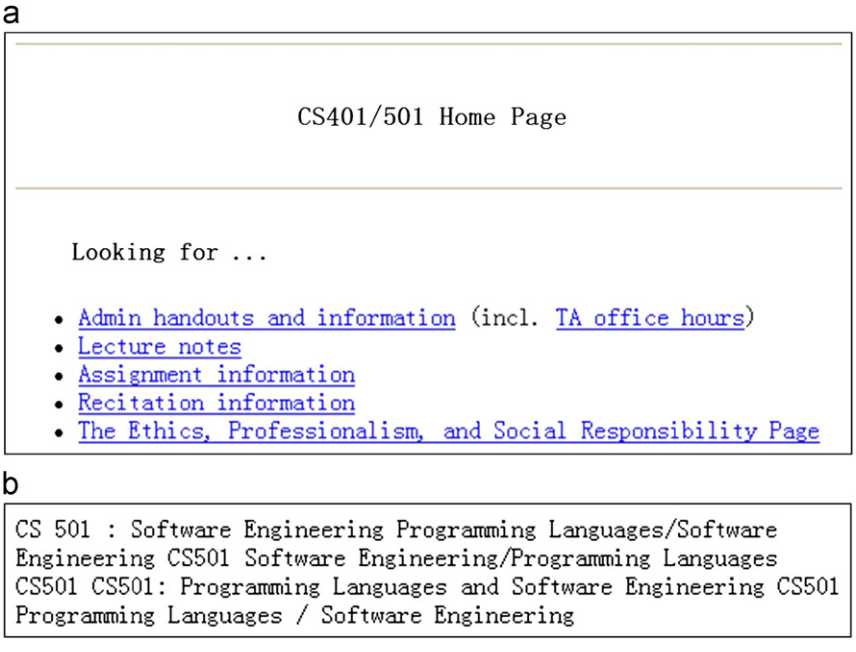

Fig. 1. Information related to a course web page in database: (a) words on the web page and (b) words in links.

is evaluated on the test fold. We report the average accuracy across the whole CV procedure. For the multi-class image classification task, the evaluation procedure is basically the same except that the original labeled examples contain one example from each category.

On all the experiments we compare ALESLE to our baseline, the general preliminary approach for active learning when only very few labeled examples are available, which constitutes of randomly selecting examples for labeling. In addition, we also adopt the ratio margin (RatioM for short) method [25] for selective sampling as another comparison on the first two data sets involving binary classification. RatioM is a support vector machine (SVM) active learning method suitable for binary classification. It first obtains the classifier margin m^+ and m^-, respectively, assuming an unlabeled example is a positive or negative example. Then RatioM chooses to query the example whose $\min(m^-/m^+, m^+/m^-)$ is largest. Being a single-view algorithm, in our experiments RatioM combines all features from different views to train linear SVMs and calculate margins.

4.1. Text classification

In this experiment, we consider the problem of classifying web pages. The data set consists of 1051 two-view web pages collected from the computer science department web sites at four U.S. universities: Cornell, University of Washington, University of Wisconsin, and University of Texas [3]. The task is to predict whether a web page is a course home page (see Fig. 1 as an example) or not. Within the data set there are a total of 230 course home pages (positive examples). The first view of the data is the words appearing on the web page itself, whereas the second view is the underlined words in all links pointing to the web page from other pages. We preprocesses each view by removing stop words, punctuation and numbers and Porter's stemming is applied on the text [19]. In addition, words that occur in five or fewer documents are ignored. This results in a 2332 and 87-dimensional vectors in space $\mathcal{V}^1$ and $\mathcal{V}^2$, respectively. Finally, document vectors are normalized to TFIDF (the product of term frequency and inverse document frequency) features [20].

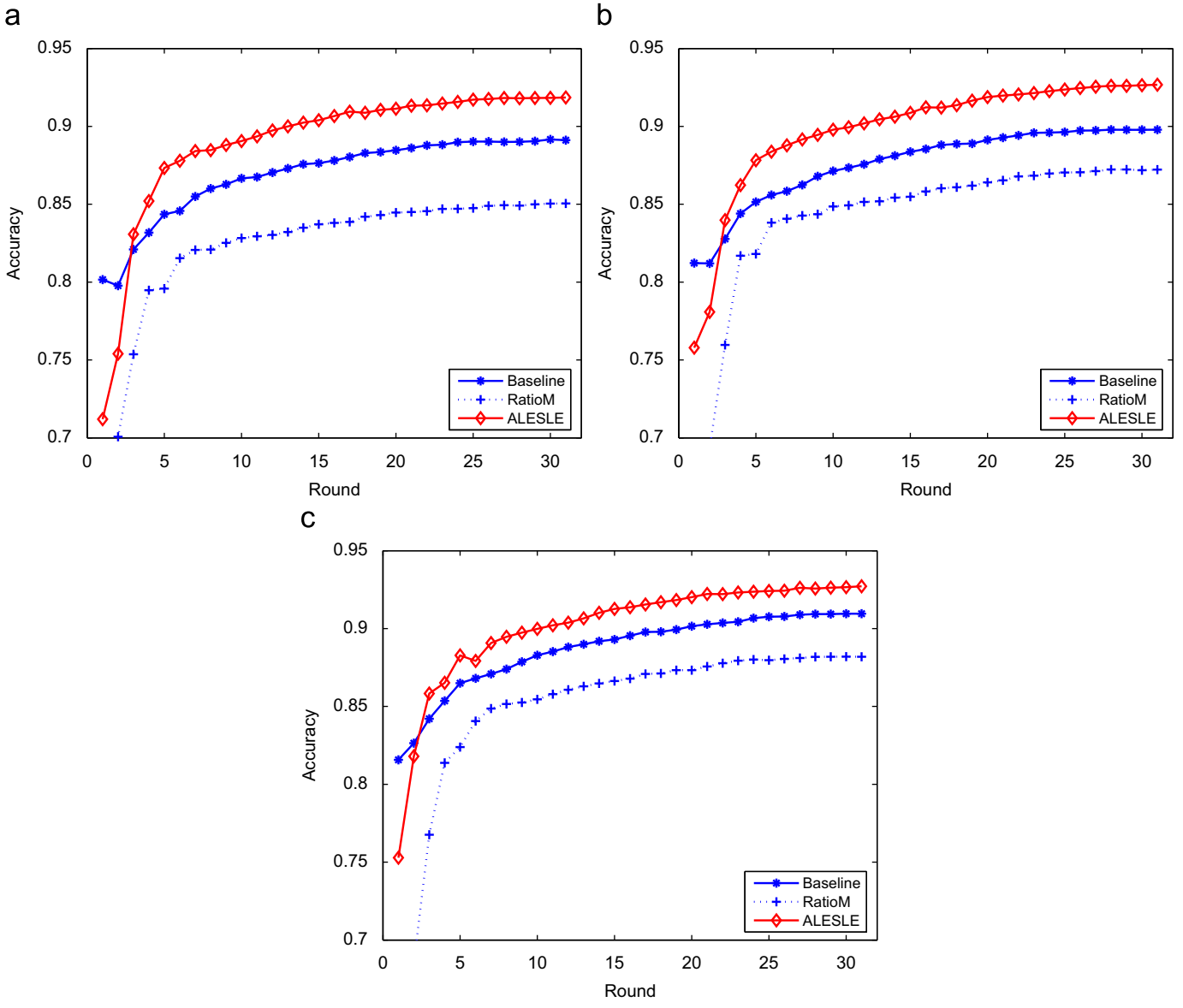

Fig. 2. Text classification performance with active learning, respectively, launched by ALESLE, RatioM and the baseline. The number of examples selected for manual labeling varies in {10, 15, 20}. (a) 10, (b) 15 and (c) 20.

We use ALESLE to select 10, 15, and 20 examples, respectively, for manual labeling. Then naive co-testing continues to make queries for 30 rounds, and after each learning episode 10 unlabeled examples are presented for labeling. On each view, linear SVM learners are employed. Besides, to evaluate the performance of active learning, after each round a single SVM classifier is trained using both views to predict the test data.

The average prediction accuracies obtained by CV for the ALESLE approach, RatioM and the baseline are shown in Fig. 2. We find that after several rounds of active learning, the performance of naive

```
...
<A href=`http://www.corp.com/sales.html`>
    Our sponsor: <IMG src=`http://www.corp.com/ads/thead.gif'
                    alt=`click here` height=`40` width=`200`></A>
...
<A href=`contact.html`>
    Contact us: <IMG src=`images/contact.gif'
                    alt=`contact info` height=`50` width=`40`></A>
```

Fig. 3. An HTML file example containing ads and non-ads image instances [15].

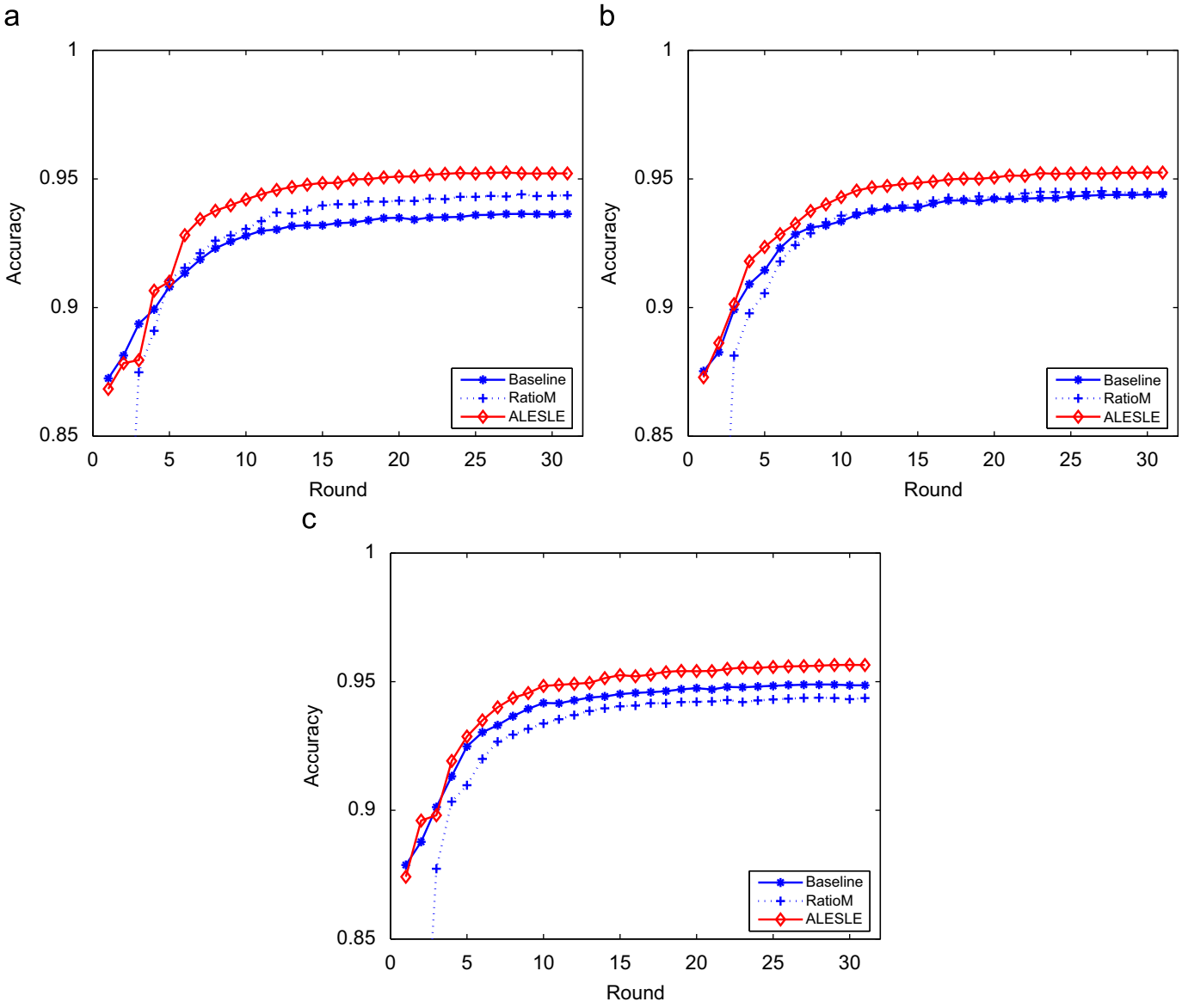

Fig. 4. Web advertisement image classification performance with active learning, respectively, launched by ALESLE, RatioM and the baseline. The number of examples selected for manual labeling varies in {10, 15, 20}. (a) 10, (b) 15 and (c) 20.

co-testing with initial examples selected by ALESLE is consistently better than that by the general random selection manner (baseline). Basically for the same number of rounds/queries, the active learning method launched by ALESLE obtains higher classification accuracies than that launched by our baseline. In other words, active learning following ALESLE can reach the same classification accuracy with

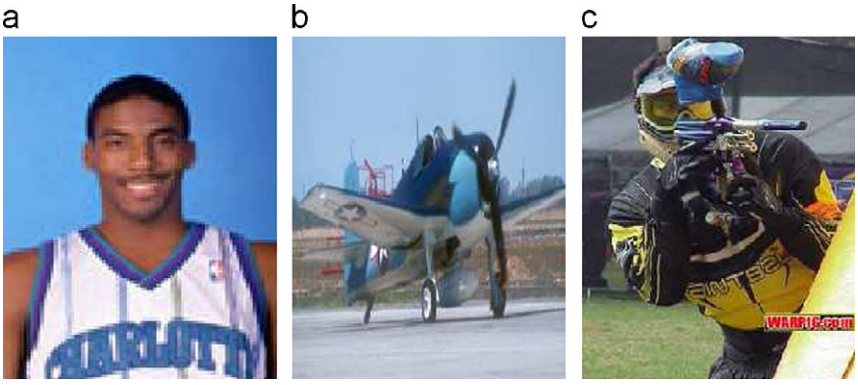

Fig. 5. Image examples in database: (a) sports, (b) aviation and (c) paintball.

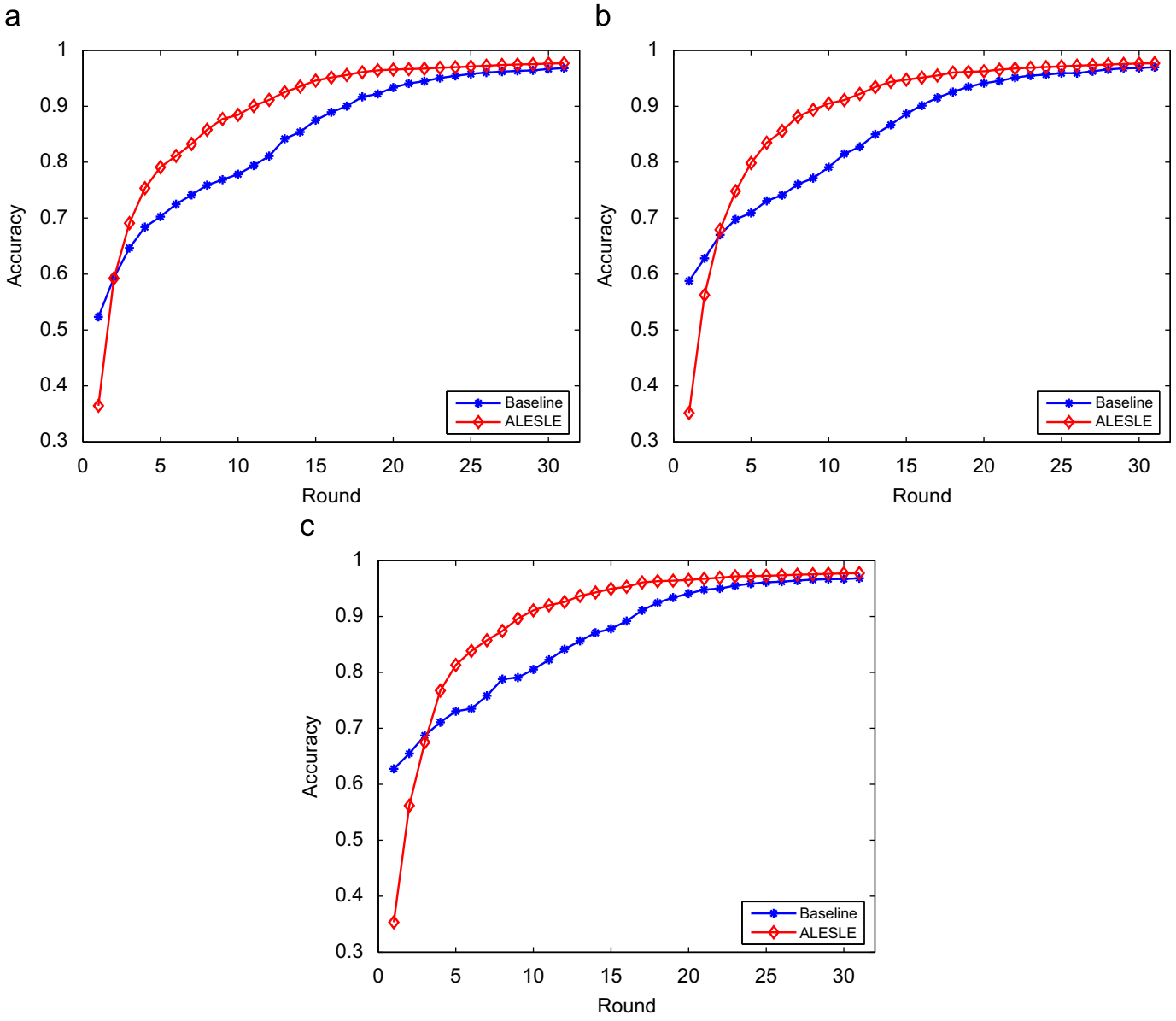

Fig. 6. Multi-class image classification performance with active learning, respectively, launched by ALESLE and the baseline. The number of examples selected for manual labeling varies in {10, 15, 20}. (a) 10, (b) 15 and (c) 20.

fewer manually labeled examples than that following the baseline. We are able to observe a performance decrease in ALESLE during the first two or three rounds. This may due to a mismatch between the distribution of the labeled data and the total distribution. Although we are able to observe that phenomenon as ALESLE can select genuinely informative examples it soon outperforms our baseline after several rounds of queries. The performance of RatioM is not as good as random selection, which is also well supported by results in [25]. That is, when the number of unlabeled examples chosen for manual labeling is not large, RatioM has no guarantee to be better than random selection.

We also tried kernel CCA [8,10] to replace the (linear) CCA in ALESLE, and found that for the current data set kernel CCA is largely inferior to CCA. We speculate that this is due to the high dimensionality of instance spaces which caused the performance degeneration of kernel CCA in ALESLE. Although Zhou et al. [27] applied kernel CCA to the same data set and obtained good semi-supervised learning results; however, the dimensionalities of their instances spaces after preprocessing are much lower, respectively, 66 and 5 for the two views. Considering this factor and the simplicity of linear methods, for the following experiments, we use CCA in ALESLE.

4.2. Advertisement removal

Advertisement images that are embedded in web pages can increase users' browsing time and distract their attentions. In this experiment we consider the problem of classifying web images into ads and non-ads [15], so that the images, which are classified as ads are then removed before the corresponding web pages are rendered to users.

The data set consists of 3279 examples with 459 of them being ads. Fig. 3 gives an example of ads and non-ads images. We use 1554 binary attributes (weights of text terms related to an image using Boolean model) whose values can be 0 and 1 for classification. These attributes are naturally divided into two views: $\mathcal{V}^1$ describes the image itself (terms in the image's URL, caption and alt text) and $\mathcal{V}^2$ contains features from other information (terms in the page and destination URLs). As a result, $\mathcal{V}^1$ and $\mathcal{V}^2$, respectively, have 587 and 967 features.

We follow the same experimental setting as in Section 4.1. Fig. 4 gives the prediction accuracies obtained by CV for ALESLE, RatioM and our baseline. For ALESLE and the baseline, curves in this figure have quite similar characteristics with that in Fig. 2. These curves show that with the same number of selected queries ALESLE can have higher performance than the general approach of random selection. In others words, to achieve a fixed accuracy ALESLE usually requires much fewer queries. As shown in Fig. 4, on this data set the performance of RatioM is better than random selection only when the number of examples selected for manual labeling is 10.

4.3. Multi-class image classification

In this case study, we try to apply ALESLE to multi-class image classification, an important problem in computer vision. A multimedia web image-text database is used, which includes examples of three classes: Sports, Aviation, and Paintball (Fig. 5 shows three examples of images) [14]. The three categories are represented as $z \in \{1,2,3\}$. Following preprocessing, $\mathcal{V}^1$ consists of 960-dimensional image visual features (image HSV color and Gabor texture), while $\mathcal{V}^2$ is composed of 3522-dimensional term frequency features appearing in the attached text.

The same experimental setting as the above two experiments is adopted. Fig. 6 gives the prediction accuracies of naive co-testing active learning, respectively, launched by ALESLE and our baseline. Tendencies reflected by the curves in this figure are similar with those in Figs. 2 and 4. These curves show that to achieve a fixed accuracy active learning launched by ALESLE can largely reduce the number of manually labeled examples. The superiority of ALESLE over the general method of randomly selecting examples to label is indicated.

Fig. 6 also suggests that with enough rounds of queries, active learning applied to the baseline approach can achieve the same performance as with ALESLE. This is straightforward to perceive if all the unlabeled examples in the training sets were labeled. But using, or labeling, all the data would oppose the main aim of active learning, which is to achieve as good a performance using a few labeled examples as possible.

5. Conclusion

In this paper we introduce ALESLE, a method for active learning with extremely sparse labeled examples. The method works under the assumption that there exists multiple views that are sufficient for correct classification. CCA between these views is used to calculate the similarities between unlabeled examples and the original labeled examples. Based on these similarities, ALESLE uses its criterion to select informative queries for manual labeling. This allows for the subsequent application of general active learning methods, such as naive co-testing, as the number of high-informative labeled examples is increased. To the best of our knowledge this is the first proposed general method for active learning with extremely limited labeled data.

ALESLE can be applied to a number of real-world scenarios, such as those described in Section 1. In this paper, empirical results on text classification, advertisement removal, and multi-class image classification tasks demonstrate the superiority of ALESLE over RatioM and the baseline approach. Experimental results have indicated that a similar number of selected queries, using active learning with ALESLE can achieve a higher performance than by using RatioM and the baseline approach of random selection. In other words, to achieve a fixed accuracy using ALESLE requires fewer queries which can greatly reduce the 'cost' of labeling.

For future work we hope to focus on the theoretical analysis of ALESLE and explore its performance with the assumption that existing views are no longer sufficient. Furthermore, we aim to explore other applications of the proposed method, for example, using the combination of ALESLE with relevance feedback in content-based image retrieval.

Acknowledgements

The authors appreciate very much the valuable comments given by the anonymous reviewers to improve this paper. This work was supported in part by the National Natural Science Foundation of China under Projects 60703005 and 61075005, and by Shanghai Educational Development Foundation under Project 2007CG30. David R. Hardoon would like to acknowledge financial support from the EPSRC Project Le Strum,[2] EP-D063612-1 and from the EU Project PinView,[3] FP7-216529.

References

[1] D. Angluin, Queries and concept learning, Machine Learning 2 (1988) 319–342.

[2] http://www.lestrum.org
[3] http://www.pineview.eu

[2] M.F. Balcan, A. Blum, K. Yang, Co-training and expansion: towards bridging theory and practice, Advances in Neural Information Processing Systems 17 (2005) 89–96.
[3] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.
[4] O. Chapelle, B. Schölkopf, Z. Zien, Semi-Supervised Learning, MIT Press, Cambridge, MA, 2006.
[5] D. Cohn, Z. Ghahramani, M. Jordan, Active learning with statistical models, Journal of Artificial Intelligence Research 4 (1996) 129–145.
[6] Y. Freund, H. Seung, E. Shamir, N. Tishby, Selective sampling using the query by committee algorithm, Machine Learning 28 (1997) 133–168.
[7] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function with application in pattern recognition, IEEE Transactions on Information Theory 21 (1975) 32–40.
[8] C. Fyfe, P.L. Lai, Kernel and nonlinear canonical correlation analysis, International Journal of Neural Systems 10 (2001) 365–377.
[9] C. Fyfe, G. Leen, P.L. Lai, Gaussian processes for canonical correlation analysis, Neurocomputing 71 (2008) 3077–3088.
[10] D. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Computation 16 (2004) 2639–2664.
[11] S. Hoi, R. Jin, J. Zhu, M. Lyu, Batch mode active learning and its application to medical image classification, in: Proceedings of 23th International Conference on Machine Learning, 2006, pp. 417–424.
[12] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.
[13] R. Jones, R. Ghani, T. Mitchell, E. Riloff, Active learning for information extraction with multiple view feature sets, in: ECML Workshop on Adaptive Text Extraction and Mining, 2003.
[14] T. Kolenda, L. Hansen, J. Larsen, O. Winther, Independent component analysis for understanding multimedia content, in: Proceedings of IEEE Workshop on Neural Networks for Signal Processing, 2002, pp. 757–766.
[15] N. Kushmerick, Learning to remove internet advertisements, in: Proceedings of Third International Conference on Autonomous Agents, 1999, pp. 175–181.
[16] D. Lewis, W. Gale, A sequential algorithm for training text classifiers, in: Proceedings of 17th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 3–12.
[17] T. Mitchell, Generalization as search, Artificial Intelligence 28 (1982) 203–226.
[18] I. Muslea, S. Minton, C.A. Knoblock, Selective sampling with redundant views, in: Proceedings of 17th National Conference on Artificial Intelligence, 2000, pp. 621–626.
[19] M.F. Porter, An algorithm for suffix stripping, Program 14 (1980) 130–137.
[20] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management 24 (1988) 513–523.
[21] H. Seung, M. Opper, H. Sompolinsky, Query by committee, in: Proceedings of ACM Workshop on Computational Learning Theory, 1992, pp. 287–294.
[22] S. Sun, C. Zhang, N. Lu, F. Xiao, A semi-supervised classification method based on transduction of labeled data, in: Proceedings of IEEE Conference on Cybernetics and Intelligent Systems, 2004, pp. 1128–1132.
[23] S. Sun, High reliable multi-view semi-supervised learning with extremely sparse labeled data, in: Proceedings of Eighth International Conference on Hybrid Intelligent Systems, 2008, pp. 935–938.
[24] S. Tong, Active learning: theory and applications, Ph.D. Thesis, Stanford University, Stanford, CA, 2001.
[25] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research 2 (2001) 45–66.
[26] Y. Zhou, S. Goldman, Democratic co-learning, in: Proceedings of 16th IEEE International Conference on Tools with Artificial Intelligence, 2004, pp. 594–602.
[27] Z. Zhou, D. Zhan, Q. Yang, Semi-supervised learning with very few labeled training examples, in: Proceedings of 22nd AAAI Conference on Artificial Intelligence, 2007, pp. 675–680.

Shiliang Sun received the B.E. degree with honors in automatic control from the Department of Automatic Control, Beijing University of Aeronautics and Astronautics in 2002, and the Ph.D. degree with honors in pattern recognition and intelligent systems from the State Key Laboratory of Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing, China, in 2007. In 2004, he was entitled Microsoft Fellow. Currently, he is an associate professor at the Department of Computer Science and Technology and the founding director of the Pattern Recognition and Machine Learning Research Group, East China Normal University. From 2009, he has been a visiting researcher at the Department of Computer Science, University College London, working within the Centre for Computational Statistics and Machine Learning. He is a member of the IEEE, a member of the PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) network of excellence, and on the editorial boards of multiple international journals. His research interests include machine learning, pattern recognition, computer vision, brain–computer interfaces and intelligent transportation systems.

David R. Hardoon is a research fellow at the Institute for Infocomm Research (I^2R) and a honorary senior research associate at the Centre for Computational Statistics and Machine Learning at University College, London. He is currently working on projects that are focused on learning the structure of music, medical analysis, multilingual and multi-modal integration. He has a keen interest in multi-view learning, kernel methods, regression, and sparsity. He has previously worked on various research projects in the fields of Taxonomy, Image analysis, classification and content-based retrieval systems. David received his first class B.Sc. Hons. in Computer Science with Artificial Intelligence from the Royal Holloway, University of London in 2002 and his Ph.D. in Computer Science in the field of Machine Learning from the University of Southampton in 2006. He has also received the Ph.D. PASCAL label award from his active participation in the PASCAL network. More information can be found on http://www.davidroihardoon.com.