Ensemble Learning with Active Data Selection for Semi-Supervised Pattern Classification

Shihai Wang and Ke Chen

Abstract—Unlike traditional pattern classification, semi-supervised learning provides a novel technique to make use of both labeled and unlabeled data for improving the performance of classification. In general, there are two critical issues for semi-supervised learning of discriminative classifiers; i.e., how to create an initial classifier of a good generalization capability with the limited labeled data and the how to make an effective use of unlabeled data without degradation of the established classifier. To tackle two aforementioned problems, we propose an ensemble learning approach based on a recent active data selection strategy [1], where ensemble learning would yield good generalization and active data selection tends to choose the unlabeled data more likely resulting in an improvement during semi-supervised learning. By using an ensemble of $K$-NN classifiers, we demonstrate the effectiveness of our approach on a synthetic data classification and a facial expression recognition task.

I. INTRODUCTION

Pattern classification is a ubiquitous task in the real world and its effective techniques are demanded from ordinary data processing to many AI areas. Traditional pattern classification techniques have been developed in the supervised learning paradigm where a number of data are first collected and labeled by experienced human annotator and then those labeled data are used to train a classifier for fulfilling the task. Apparently the annotation of data for labeling demands a great deal of human effort, which results in a hard, expensive and time-consuming process prior to learning. On the other hand, a large amount of unlabeled data can be collected easily. As a result, semi-supervised learning provides a novel technique to employ both labeled and unlabeled data for improving the performance of pattern classification with less human effort.

In the past decade, many semi-supervised learning approaches have been developed [2],[3]. In general, these approaches can be classified into two categories in terms of the structure of a classifier used; i.e., generative model and discriminative model. From a viewpoint of statistics, the discriminative model is viewed as a posterior probability estimator for a specified label given an input while the generative model specifies a joint probability model between a specific label and an input. Due to the nature of joint probability, the generative model can be easily modified to use both labeled and unlabeled data for training [2]-[6]. However, the classification performance of such generative models is often less satisfactory due to their weak discriminative capability. For the discriminative model in the supervised learning paradigm, the distribution of input is not considered and hence it is hard to include the information regarding the input distribution for semi-supervised learning. Instead efforts are made by exploiting other information sources [7] and using heuristics [8]. In such discriminative models, there are two critical issues for semi-supervised learning; i.e., how to create an initial classifier of a good generalization capability with the limited labeled data and the how to make use of unlabeled data properly without degradation of the established classifier. Due to the lack of the input distribution information, it is often difficult to deal with the second issue. Ensemble learning provides a reliable way to create a classifier ensemble of good generalization. As a result, boosting algorithms have been adopted for semi-supervised learning [9],[10]. The use of a boosting algorithm leads to a good generalization by indirectly or implicitly considering the input distribution information via its re-sampling mechanisms. As pointed out in [11], however, there are few ensemble learning methods developed for semi-supervised learning in general.

In this paper, we propose an ensemble learning approach based on a recent active data selection strategy [1] for semi-supervised pattern classification. The strength of this recently proposed active data selection strategy directly or explicitly takes not only confidence scores but also the input distribution of unlabelled data into consideration. Taking advantage of this strength, we create an ensemble of discriminative classifiers with bootstrap re-sampling techniques or different representations of raw data. During semi-supervised learning, unlabelled data are actively selected by the performance-distribution driven strategy [1] to train each component classifier based on the consensus of other component classifiers in the ensemble. Thus we expect such an idea would effectively deal with two issues in question. By using an ensemble of $K$-nearest neighbor classifiers [12], we demonstrate the effectiveness of our approach on a synthetic data classification and a facial expression recognition tasks.

The rest of this paper is organized as follows. Sect. II reviews the active selection strategy and presents our ensemble semi-supervised learning approach. Sect. III describes simulations and reports results. The last section discusses the related issues and draws a conclusion.
II. DATA-SELECTION BASED ENSEMBLE LEARNING

In this section, we first describe a modified version of the recent active data selection strategy [1] used for an ensemble of discriminative classifiers. Then we propose an ensemble learning approach to annotate unlabeled data for semi-supervised learning. Finally we present the pseudo code of our data-selection based ensemble semi-supervised learning algorithm.

A. Active Data Selection Strategy

As argued in [1], many existing semi-supervised learning algorithms select unlabeled data based on only their confidence annotation produced by a classifier trained on the labeled data. A consequence of selecting only unlabeled data of high confidence scores for re-training the classifier simply reinforces what the model has already encoded and is unlikely to reduce the estimation bias by scarcity of labeled data or an inaccurate model assumption, which leads to an erroneous estimate of the true input distribution, even if the assigned labels are correct, and therefore may not improve the performance.

In order to tackle this problem, Zhang and Rudnicky recently proposed a performance-driven active data selection strategy [1] by explicitly considering the input distribution along with confidence scores. Their basis idea is summarized as follows: a) partitioning the input space into several inherent subspaces by clustering analysis and using confidence scores to generate a set of ordered data groupings, b) following the order, each grouping of unlabeled data will be incorporated to train the learner and its performance is evaluated, c) accepting the grouping of unlabeled data only if the classifier trained on this data grouping outperforms its older version, and d) repeating b) and c) until all available unlabeled data have been examined.

Suppose there are $|L|$ labeled data, $\{(x_i,y_i)\}_{i=1}^{l}$, where $y_i \in \{c_{a_i}\}_{a_i=1}^{M}$ and $|U|$ unlabeled data, $\{x_{bi}\}_{bi=1}^{n}$, available. For a discriminative probabilistic classifier $h$, its output, $P_h(c \mid x)$, is the estimate of posterior probability of class $c$ given the input $x$. A confidence score measure based on $h$ can be defined by the negative entropy [8] as follows:

$$\text{neg}_{\text{entropy}}(x \mid h) = \sum_{m=1}^{M} P_h(c_m \mid x) \log[P_h(c_m \mid x)]$$  

(1)

This measure reflects the uncertainty that we make a decision based on the estimate of posterior probabilities. Assume that the unlabeled data are partitioned into $N$ equal-sized groupings stored in bins. The data partitioning algorithm [1] for ensemble learning is as follows:

**Data Partitioning Algorithm**

- Initialize each with empty set for classifier $h$:
  $$B_{i,0} = \{b_{i,0}^1, b_{i,0}^2, \ldots, b_{i,0}^{N} \} \quad \text{null} (n=1,2,\ldots,N);$$

- Estimate posterior probability, $P_h(c_m \mid x) (m=1,\ldots,M)$, for each unlabeled datum based on the classifier ensemble $H$, described in Sect. II.B, and compute its confidence score, $\text{neg}_{\text{entropy}}(x \mid H)$, using (1);

- Split input space into $K$ subspaces $D_1, D_2, \ldots, D_k$ with a clustering algorithm;

- For each subspace $D_k$ ($1 \leq k \leq K$):
  - Sort the unlabeled datum $x$ that $x \in D_k$ according to its confidence score from high to low;
  - Add each unlabeled datum $x$ that $x \in D_k$ to one of the $N$ bins in the way that $b_{i,n}^k (1 \leq n \leq N)$ accepts the data which confidence scores are with the range from top $n-1 \frac{1}{N} \%$ to $n \frac{1}{N} \%$;

A fundamental principle for selecting unlabeled data is that their inclusion would not degrade the performance of the established classifier. Due to a lack of information on the ground truth of unlabelled data, a pseudo performance measure is required to evaluate the performance of a re-trained classifier on a grouping of unlabeled data produced by the above partitioning algorithm. By the MAP principle, a pseudo-class label is assigned to each unlabeled datum by the current classifier $h$; i.e., $y_i = \arg \max_{y \in \{c \mid x\}} P_h(c \mid x)$. Thus a universal classification capability measure applicable to both labeled and unlabeled data is defined as follows:

$$d_i(y_i \mid x) = \log P_h(y_i \mid x) - \log \left( \frac{1}{M-1} \sum_{y \neq y_i} P_h(y \mid x) \right)$$

(2)

Here, $y_i$ is the label of $x_i$ if it is a labeled datum and the pseudo-class label otherwise. Note that the classifier capability measure in (2) is especially for a discriminative classifier, which is different from that defined in [1].

The normalized overall performance of classifier $h$ on a given data set is evaluated by

$$f(h) = \frac{1}{J} \sum_{j=1}^{J} \exp[ -\beta d_i(y_j \mid x_j) + \theta ]$$

(3)

where $\beta$ and $\theta$ control the slope and transition point of the sigmoid function. If $J=|L|$, the measure gives the performance on the labeled data, while it reflects the pseudo performance on the mixed set of labeled and unlabeled data if $J=|L|+|U|$. Thus, the use of (3) can achieve the pseudo performance of a classifier as a grouping of unlabeled data are used for training and therefore can be employed to select unlabeled data by checking whether adding the grouping of unlabeled data improves the established classifier or not.

B. Classifier Ensemble Generation

Co-training presents an effective method by using two classifiers for semi-supervised learning [7]. Nevertheless it demands the availability of two independent and redundant feature sets. Motivated by the co-training idea, we propose an ensemble learning approach for semi-supervised learning.
In general, there are two circumstances for a pattern classification task. One is that the nature of the task makes different redundant representations\(^1\) available, and the other is no different representations available. For the former case, our earlier studies show the usefulness of combining multiple classifiers trained on different representations [13]-[15]. For the latter case, we can use the bootstrap re-sampling techniques [16] to create different data sets and then train an ensemble of classifiers on them. As a result, we are always able to create an ensemble of classifiers for robust learning regardless of the nature of a pattern classification task.

For annotating an unlabeled datum to re-train a component classifier, we extend the co-training idea to an ensemble of classifiers; i.e., we combine all other established component classifiers to reach a consensus to yield a pseudo-class label for the unlabeled datum. For the component classifier \(h_i\), the pseudo-class label of \(x\), \(H_i(x)\), is assigned by the ensemble of all other component classifiers, \(H_{-i}\), as

\[
H_i(x) = \arg \max \sum_{j \in \{1,\ldots,I\} \setminus \{i\}} \alpha_j h_j(x) | \bullet), \tag{4}
\]

where \(I\) is the number of classifiers in the ensemble, \(h_j(x) | \bullet)\) is the output of classifier \(h_j\) (the posterior probability if \(h_j\) is a probabilistic classifier), based on a representation of \(x\) for the case of different representations or itself, and the weight, \(\alpha_p\), is defined based on the error rate of the current classifier \(h_j\), \(\epsilon_j\), measured on the labeled data set:

\[
\alpha_p = \frac{\exp(-\epsilon_p)}{\sum_{i \in \{1,\ldots,I\}} \exp(-\epsilon_i)} \tag{5}
\]

The above annotation procedure works alternately until every unlabeled datum has a pseudo-class label for each of component classifiers. Then each component classifier is trained based on its own training set generated by the active data selection strategy in last subsection. When no unlabeled data are left, the classifier ensemble will be established.

C. Data-selection based ensemble learning algorithm

Now we summarize our data-selection based ensemble learning algorithm for semi-supervised pattern classification as follows:

**Ensemble Semi-Supervised Learning Algorithm**

Input:
- \(L\): original labeled data set
- \(U\): unlabeled data set
- \(T\): testing data set
- \(\text{learn}(h, S)\): learning algorithm for classifier \(h\) trained on \(S\)
- \(I\): the number of re-sampled training sets or different representations extracted from raw data

Output: for a testing datum \(x \in T\),

\[
H(x) = \arg \max \sum_{i \in \{1,\ldots,I\}} \alpha_i h_i(x) | \bullet)
\]

where \(\alpha_i = \exp(-\epsilon_i) / \sum_{i=1}^I \exp(-\epsilon_i)\) determined by the error rate of all \(I\) classifiers similar to that in (5), and \(h_i(x) | \bullet)\) is the output of the classifier \(h_i\) similar to the one in (4).

\(\text{for } i = 1,\ldots,I \text{ do }\{
\begin{align*}
& \text{if no different representations available } \\
& \quad \circ \text{ create a training set } S_{i,0} \text{ based on } L \text{ using a bootstrap re-sampling technique. } \\
& \text{else } \\
& \quad \circ \text{ extract features to form a representation set } S_{i,0} \text{ for } h_{i,0} \text{ by the data partitioning algorithm in Sect. II.A with } H_{i,0} \text{ in (4). } \\
& \end{align*}
\}
\text{end_of_for}

\text{for } i = 1,\ldots,I \text{ do } \{
\begin{align*}
& \text{partition } U \text{ into } N \text{ bins } B_{i,0} = \{b_{i,0}^1, b_{i,0}^2, \ldots, b_{i,0}^N\} \text{ for } h_{i,0} \text{ by } \\
& \text{the data partitioning algorithm in Sect. II.A with } H_{i,0} \text{ in (4). } \\
& \text{end_of_for}
\}
\text{set } k = 0; \text{ repeat until } B_{i,k} = \text{null } (i = 1,\ldots,I) \{
\begin{align*}
& \text{for } i = 1,\ldots,I \text{ do } \{
& \text{if } B_{i,k} \text{ is not null } \\
& \quad \text{\quad \circ } \text{ for } b \in B_{i,k} \text{ do } \\
& \quad \quad \circ h_{i,b,k}^{k+1} = \text{learn}(h_{i,k} S_{i,k} \cup b) ; \\
& \quad \quad \circ \text{ evaluate the performance } f(h_{i,k}^{k+1}) \text{ with (3); } \\
& \quad \text{\quad } \}\text{end_of_for}
& \text{\quad \circ } b^* = \arg \max_{b \in B_{i,k}} f(h_{i,b,k}^{k+1}) ; \\
& \text{\quad } S_{i,k+1} = S_{i,k} \cup b^* ; \\
& \text{\quad } h_{i,k+1} = h_{i,b^*,k}^{k+1} ; \\
& \text{\quad } B_{i,k+1} = B_{i,k} - b^* ; \\
& \text{\quad } \text{Re-label all unlabeled data in } B_{i,k+1} \text{ by the classifier } \\
& \text{\quad } \text{ensemble } H_{i,k+1} \text{ with (4). } \\
& \quad \text{\quad } k \leftarrow k + 1; \\
& \text{\quad } \}\text{end_of_for}
& \}\text{end_of_repeat}
\}
\]

\(\text{\textsuperscript{1}}\) Here different representations are not assumed to be independent and can be employed as long as they are not completely correlated.
III. SIMULATIONS

In this section, we present our experimental methods and report simulation results. In order to demonstrate the performance on two different circumstances of pattern classification tasks, we apply our approach to a 4-category synthetic data classification task without the use of any representation and a facial expression recognition task where three different representations are available.

A. Experiment Description

In our experiment, we employ a $K$-nearest neighbor ($K$-NN) classifier [12] as a base classifier. The use of the nearest neighbor rule with the Mahalanobis distance results in the posterior probability estimate for each class given an input datum; i.e., $P(c_m \mid x) (m = 1, \ldots, M)$ for an $M$-category classification [12]. For robust learning and better generalization, we further use the multiclass boosting algorithm [17] to generate a strong classifier used as a component classifier in our ensemble semi-supervised learning algorithm presented in Sect. II.C. Note that although such a strong classifier is an ensemble of $K$-NN classifier in nature, we still name it single strong classifier hereinafter to avoid the confusion with the use of multiple such strong classifiers to form a classifier ensemble in our semi-supervised ensemble learning algorithm. Due to its nature of the multiclass boosting algorithm [17], the output of such a strong classifier is also the posterior probability estimate.

In all the simulations, we use a 3-NN classifier as the base learner. For creating each single strong classifier, 20 boosting rounds are taken by the multiclass boosting algorithm [17]. For the active unlabeled data selection, we use the $K$-mean algorithm ($K=4$) and three bins in the data partitioning algorithm described in Sect. II.A for both a single strong classifier with the active data selection strategy [1] and our ensemble semi-supervised learning algorithm where an ensemble of strong classifiers are employed.

For each task, we report results of supervised learning of a single strong classifier trained on the labeled data set only, semi-supervised learning of a single strong classifier with the active data selection strategy [1] and our data-selection based ensemble semi-supervised learning presented in Sect. II.C. Ten trials have been done for each experiment and the averaging results plus their deviation are reported here.

B. Synthetic Data Classification

A four-category synthetic data set of 800 data points is generated based on four Gaussian distributions and there are 200 points for each category, as illustrated in Fig. 1. We use this linearly non-separable data set for evaluating the performance of different algorithms.

In each trial of our simulations, we randomly divide the data set into three subsets; i.e., labeled data set, unlabeled data set and testing data set. For thoroughly investigating the performance of semi-supervised learning, we adopt different labeling rate to form labeled and unlabeled data set from the remaining data after we first randomly select 20% data points as the testing data set ($T$). From the remaining data (i.e. 80% of the whole data set after excluding 20% testing data), we randomly choose 10%, 20%, 30%, 40% and 50% of them to form a labeled data set ($L$), respectively, and the rest of data would constitute its corresponding unlabeled data set ($U$) where their label information is reserved without use during semi-supervised learning.

Since we directly cope with the 2-D data classification task without feature extraction, we employ the bagging algorithm [16] for re-sampling of labeled data so that five re-sampled labeled data sets are generated; i.e., $I=5$ in our ensemble semi-supervised learning algorithm in Sect. II.C; in other words, the classifier ensemble used in our algorithm consists of five strong classifiers initially trained on re-sampled labeled data sets, respectively.

<table>
<thead>
<tr>
<th>Labeling Rate</th>
<th>Supervised Learning (labeled data)</th>
<th>Semi-supervised Learning (single strong classifier)</th>
<th>Semi-supervised Learning (ensemble)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>75.00±3.61</td>
<td>75.31±3.66</td>
<td>75.78±4.37</td>
</tr>
<tr>
<td>20</td>
<td>77.06±4.45</td>
<td>76.25±4.30</td>
<td>78.31±4.66</td>
</tr>
<tr>
<td>30</td>
<td>77.58±3.77</td>
<td>78.28±2.71</td>
<td>78.59±3.76</td>
</tr>
<tr>
<td>40</td>
<td>78.26±3.88</td>
<td>77.06±4.22</td>
<td>78.68±3.39</td>
</tr>
<tr>
<td>50</td>
<td>78.20±2.22</td>
<td>78.44±2.30</td>
<td>78.54±2.67</td>
</tr>
</tbody>
</table>

Table I shows the performance of supervised learning of a single strong classifier trained on labeled data set only, semi-supervised learning of a single strong classifier with the active data selection strategy [1] and our data-selection based ensemble semi-supervised learning algorithm. From Table I, it is evident that the use of our ensemble semi-supervised learning algorithm always yields an improvement no matter how many labeled data are used in creating initial five
component classifiers. In particular, a considerable recognition rate gain is obtained when only 20% and 30% of labeled data are used, respectively, for such a data set of a high optimal Bayesian classification error. In contrast, the use of a single strong classifier with the active data selection strategy leads to the better performance as 10%, 30% and 50% labeled data are used but its performance becomes worse than the supervised learning as 20% and 40% labeled data are used.

From the simulations above, we conclude that the use of ensemble learning yields the better generalization and robust learning in semi-supervised learning. Here we emphasize that even though no independent and redundant feature sets are available we can still use a bootstrap re-sampling technique to create a classifier ensemble for robust semi-supervised learning.

C. Facial Expression Recognition

Automatic facial expression recognition is a task to use facial images of an individual to identify his/her emotional state at the moment that the pictures were taken. Emotional state detection/recognition is one of the most important tasks in the human-computer interaction (HCI). Due to the accessibility of facial images, automatic facial expression recognition becomes a prior domain to be studied in HCI. While a large amount of facial images are available easily, the annotation of a facial image with a proper emotional state is often extremely difficult since individuals express their emotion in a rather diversified way and facial expression is also affected by a number of factors, e.g., the culture, habits and so on. Thus labeling facial expression examples would be a hard, expensive and time consuming job. The nature of automatic facial expression recognition suggests that it is a proper task where the semi-supervised learning paradigm can be applied. Therefore, we apply our ensemble semi-supervised learning algorithm to such a task to evaluate its performance further.

Japanese Female Facial Expression (JAFFE) data set [18] is a benchmark designed especially for static facial expression recognition. In this data set, 119 facial images of ten Japanese women are collected and annotated manually as four different emotional states: i.e., anger, happiness, surprise, and neutral, as illustrated in Fig. 2 for example. For facial expression recognition, most of existing methods need to extract a representation from a raw facial image. There are a number of facial image representations for such a task. In general, these representations can be divided into two categories based on their mechanisms; i.e., global and local representations. A global representation treats the image as a whole and therefore tends to encode features of a facial expression in a global way, while a local representation encodes local salient features of different facial areas to reflect the characteristics of a facial expression. Since two kinds of representations convey various facial features from different perspectives, they are likely to be complementary and redundant to some degree although there is no guarantee for them to be independent.

In our simulations, we choose three different kinds of representations for facial expression recognition; i.e., Independent Component Analysis (ICA) [19], Local Binary Patterns (LBP) [20] and 2-D Discrete Cosine Transform (DCT) [21]. ICA is a global representation that consists of a series of components decomposed from the original image by making them as independent as possible. In our experiment, 40 such components are extracted from a 200×200 facial image to form the ICA representation. To a great extent, LBP can be viewed as a local representation where local texture information of a facial image is encoded by LBP operators based on non-overlapping local image patches of a facial image. In our experiment, the size of an image patch is 40×40 pixels. Finally 2-D DCT is another global representation by transforming a facial image from its spatial domain to its frequency domain, which could encode some salient features that cannot be uncovered in spatial domain. Since the lower frequencies often carry rich information on the facial expressions [21], we employ 20×20 2-D DCT coefficients corresponding to the lower frequencies as the 2-D DCT representation used in our experiment.

![Fig. 2. Examples in JAFFE benchmark data set.](image)

In each trial of our simulations, we randomly divide the data set into three subsets to form the labeled data set (L) (27% of total data), the unlabeled data set (U) (53%) and the testing set (T) (20%). Due to the availability of three aforementioned representations, we naturally create a classifier ensemble of three strong classifiers trained on individual representations, respectively, as required by our ensemble semi-supervised learning algorithm described in
Sect. II.C. For comparison, we also employ a strong classifier trained on each individual representation for supervised learning based on L only and semi-supervised learning with the active data selection strategy [1] based on L+U.

Fig. 3 shows the performance of different learning algorithms on the JAFFE data set. It is evident from Fig. 3 that in general our ensemble semi-supervised learning on different representations outperforms the supervised learning and the semi-supervised learning with the active data selection strategy [1] on individual representations. On average, our ensemble semi-supervised system trained on different representations of both labeled and unlabeled data leads to at least 3.5% higher recognition rate than the supervised learning system trained on any individual representations of labeled data only. From Fig. 3, it is also observed that the semi-supervised learning system of a strong classifier merely yields an improvement only when the ICA representation is used but fails to work on other two representations. By comparison, we conclude that the simultaneous use of different representations results in robust learning and better generalization during semi-supervised learning, which is completely consistent with our previous argument in unsupervised [13] and supervised learning [14],[15].

While we claim that in general our ensemble semi-supervised learning algorithm outperforms others used in the simulations mentioned above, we notice that there is a larger deviation in our results among ten trials. One reason we reckon would be that the limited number of testing images (only 24 images) could cause such a performance fluctuation reflected in percentage. In our ongoing study we shall be looking for a larger facial expression benchmark data set to investigate this problem.

In summary, our simulation results in the synthetic data classification and the facial expression tasks demonstrate the usefulness of further introducing ensemble learning to semi-supervised learning via either re-sampling techniques or the simultaneous use of different representations.

IV. CONCLUSIONS

In this paper, we have presented a data-selection based ensemble learning algorithm for semi-supervised pattern classification. The use of ensemble learning leads to an improvement, which has been demonstrated by a synthetic data classification and a facial expression recognition tasks. Our preliminary work suggests that even without the use of independent feature sets like co-training ensemble learning works very well for semi-supervised learning and the exploitation of input distribution information, e.g. the active data selection strategy [1], in discriminative model for semi-supervised learning is promising for improving the performance.

Nevertheless it is worth mentioning that from a computational perspective our ensemble semi-supervised learning algorithm has a higher computation burden during training, which could be a problem as applied to a large scale problem of hundreds and thousands data points in a high dimensional space. In our ongoing work, we are exploring alternative yet efficient ensemble discriminative learning approaches with the capability of exploiting input distribution and other information sources.

ACKNOWLEDGMENT

Authors are grateful to R. Zhang for a discussion on their active data selection strategy.

REFERENCES