Expert Systems with Applications 41 (2014) 2372–2378

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# Semi-supervised learning combining co-training with active learning

Yihao Zhang [a], Junhao Wen [b,*], Xibin Wang [a], Zhuo Jiang [a]

[a] College of Computer Science, Chongqing University, Chongqing 400030, China
[b] College of Software Engineering, Chongqing University, Chongqing 400030, China

ARTICLE INFO

*Keywords:*
Semi-supervised learning
Co-training
Confidence estimation
Active learning
Informative instances

ABSTRACT

Co-training is a good paradigm of semi-supervised, which requires the data set to be described by two views of features. There are a notable characteristic shared by many co-training algorithm: the selected unlabeled instances should be predicted with high confidence, since a high confidence score usually implies that the corresponding prediction is correct. Unfortunately, it is not always able to improve the classification performance with these high confidence unlabeled instances. In this paper, a new semi-supervised learning algorithm was proposed combining the benefits of both co-training and active learning. The algorithm applies co-training to select the most reliable instances according to the two criterions of high confidence and nearest neighbor for boosting the classifier, also exploit the most informative instances with human annotation for improve the classification performance. Experiments on several UCI data sets and natural language processing task, which demonstrate our method achieves more significant improvement for sacrificing the same amount of human effort.

## 1. Introduction

Semi-supervised learning is very useful in many practical applications, which learn from both labeled data and unlabeled data and automatically exploit unlabeled data for improving the learning performance without human intervention (Chapelle, Scholkopf, & Zien, 2006; Zhu, 2008). Co-training is a well-known Semi-supervised learning paradigm started from Blum and Mitchell's seminal work (Blum and Mitchell, 1998), which is proposed for binary classification problems in which two different views are available. The standard co-training algorithm requires two sufficient and redundant views (Blum and Mitchell, 1998), that is, the attributes can be naturally partitioned into two sets, each of which is sufficient for learning and conditionally independent to the other given the class label. Co-training works in an iterative manner that two classifiers are trained separately on the different views and the predictions of either classifier on unlabeled instances are used to augment the training set of the other (Zhang & Zhou, 2011; Zhu, 2008).

There are a notable characteristic shared by many co-training algorithm: the selected unlabeled instances should be predicted with high confidence, since a high confidence score usually implies that the corresponding prediction is correct (Blum and Mitchell, 1998; Mihalcea, 2004). Unfortunately, it is not always able to improve the classification performance with these high confidence unlabeled instances. Tang et al. (2007) proposed a new strategy that updating the classifiers through co-training, which add negative instances that are close to the classifier hyper-plane such that the classifier will learn to better distinguish these instances. Following the work on standard co-training, a number of relevant methods have been developed. Wang and Zhou (2010) analyzed the co-training process and viewed it as combinative label propagation over two views. Yu, Krishnapuram, Rosales, and Rao (2011) proposed a Bayesian undirected graphical model for co-training, which can elegantly handle data samples with missing views. Sun et al. (2011) proposed an entity-based co-training algorithm, which requires no prior knowledge about the underlying class distribution which is crucial in standard co-training algorithms. Unfortunately, co-training is called for precisely when the labeled training set is small, and it is uncertain whether the standard co-training would work or not on small labeled training sets (Du et al., 2011).

In this paper, a new semi-supervised classification algorithm was proposed which combines the benefits of both co-training and active learning, and the major contributions are two-fold:

(1) Firstly, in each co-training round, a few of most reliable instances were picked out from the unlabeled data for the next round of learning, and the most reliable instances were chosen according to the two criterions of high confidence and nearest neighbor. Specifically, the contribution degree was defined as the criteria of select informative instances, which not only considering the most uncertain of instances but also considering the uncertainty difference between the instance and its nearest neighbor.

* Corresponding author. Tel.: +86 13983146919.
*E-mail addresses:* yihaozhang@cqu.edu.cn (Y. Zhang), jhwen@cqu.edu.cn (J. Wen), binxiwang@cqu.edu.cn (X. Wang), jiangzhuo1986@gmail.com (Z. Jiang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.035

(2) Secondly, the active learning uses query framework that an active learner queries the instances about which it is least certain how to label, our algorithm defined contribution degree as the selection criteria of informative instances, which achieved more significant improvement for sacrificing the same amount of human effort and worked well on small labeled training sets.

The rest of this paper is organized as follows. Section 2 reviews some issues in the co-training and active learning. After that Section 3 introduces the sketch of the algorithm and presents the algorithms details. Section 4 reports experimental results on a number of real-world datasets and further analyzes the underlying reasons for the algorithm. Finally, Section 5 concludes and indicates several issues for future work.

## 2. Issues in co-training and active learning

Co-training is a semi-supervised, multi-view algorithm that uses the initial labeled data set to learn a weak classifier in each view (Blum and Mitchell, 1998). Then each classifier is applied to the rest of unlabeled instances, and co-training detects the instances on which each classifier makes the most confident predictions. These high-confidence instances are labeled with the estimated labels and added into the labeled data set. Based on the new training set, a new classifier is repeated for several iterations. At the end, a final hypothesis is created by a voting scheme that combines the predictions of the classifiers learned in each view.

Co-training, a good paradigm of semi-supervised learning, has drawn considerable attentions and interests recently (Zhou & Li, 2010). The standard co-training assumes that the data can be described by two disjoint sets of features or views, and it works well when the two views satisfy the sufficiency and independence assumptions (Blum and Mitchell, 1998). However, these two assumptions are often not known or ensured in practice, and view splitting is unreliable under the given small labeled training sets. More commonly, most supervised data sets are described by one set of attributes (one view). To exploit the advantage of co-training, Goldman and Zhou (Goldman and Zhou, 2000) proposed an algorithm which does not exploit feature partition; the algorithm uses two different supervised learning algorithms to train the two classifiers. Zhou and Li (2005) proposed the tri-training approach, which uses three classifiers generated from bootstrap samples of the original training set. Du and Ling (2011) got the conclusions that co-training's effectiveness are mixed. That is, if two views are given, and known to satisfy the two assumptions, co-training works well; Otherwise, based on small labeled training sets, verifying the assumptions or splitting single view into two views are unreliable; thus, it is uncertain whether the standard co-training would work or not.

Active learning, a subfield of machine learning, can perform better with less training by choosing the data form which it learns. It attempts to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle (Settles, 2010) (e.g., a human annotator). In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. The key idea behind most active learning algorithms is to select the instances that are most uncertain to classify. Therefore, a key aspect of active learning is to measure the classification uncertainty of unlabeled instances. Zhu (2003) proposed a new semi-supervised learning strategy, which combines active learning and semi-supervised learning under a Gaussian random field model. Yang et al. (2009) proposed Bayesian framework to active distance metric learning by selecting those unlabeled example pairs with the greatest uncertainty in relative distance. Lughofer (2012) proposed a novel active learning strategy for data-driven classifiers, which is essential for reducing the annotation and supervision effort of operators in off-line and on-line classification systems, as operators only have to label an exquisite subset of the off-line training data. Li, Shi, and Liu (2012) proposed a joint active learning approach which combines a novel generative query strategy and the existing discriminative one, which adaptively fits the distribution difference and shows higher robustness than the ones using single strategy.

From the above analysis, co-training is an important technique for improving the predictive accuracy when labeled data are scarce. However, this algorithm is often not ensured work well in real world application. Firstly, co-training requires the data set can be splits two views, and satisfy the sufficiency and independence assumptions. In practice, those conditions are not easy to achieve. Secondly, although co-training usually selects high confidence instances that are labeled with the estimated class labels and add them to the training sets, which does not ensured these selected high confidence instances are more valuable for improving the predictive accuracy. In the paper, a semi-supervised algorithm combining co-training with active learning was proposed, which can utilize the benefits of the two algorithms and reduce the annotation effort of operators.

## 3. Combining co-training with active learning

In this section we provide a high-level description of the semi-supervised learning algorithm, and its framework can be described as Fig. 1.The algorithm (SSLCA) which combines co-training with active learning can be divided into three steps: firstly, the labeled data was split into two views in order to apply the standard two-view co-training, which learns the classifier $h_1$ and classifier $h_2$ based solely on the two views of labeled data; secondly, the unlabeled data also was split into two views for estimating their confidence using separate classifier; thirdly, the most reliable instances or informative instances were selected based on some strategy. The most informative instances were chosen according to two criterions of high confidence and nearest neighbor, and then were put into another pool for further annotation.

### 3.1. Confidence estimation methods

#### 3.1.1. Naive Bayes method

Naïve Bayes forms maximum a posteriori estimates for the class conditional probabilities for each feature from the labeled training data $D$. The prior probabilities of each class are calculated in a similar fashion by counting over instances. Define $P(c_j)$ denotes the probabilities of class $c_j$, and $|D|$ denotes the number of instances in training data:

$$P(c_j) = \frac{1 + \sum_{i=1}^{|D|} P(c_j)}{|D|}$$

Then the posteriori probabilities estimates for each instance are calculated according to the independence assumption, define $a_i$ denotes the each feature in each instance, $n$ denotes the number of the features:

$$P(c_j|a_i) \propto P(c_j)P(a_i|c_j) = P(c_j)\prod_{i=1}^{n} P(a_i|c_j)$$

#### 3.1.2. Expectation Maximization method

Expectation Maximization (EM) is an iterative statistical technique for maximum likelihood estimation in problems with

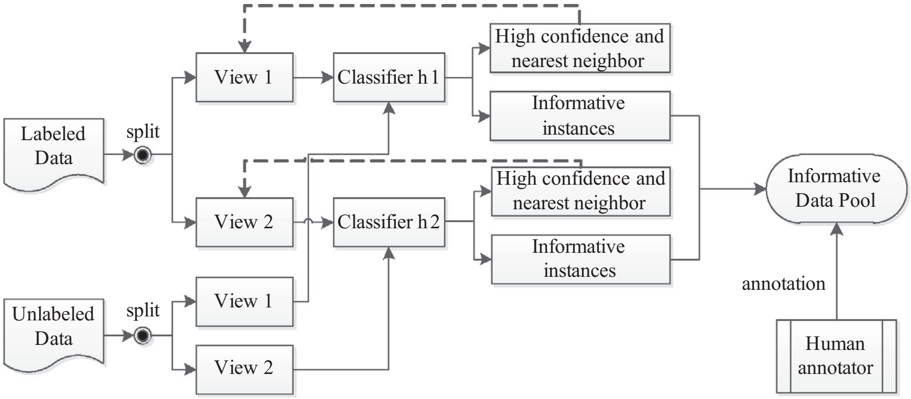

**Fig. 1.** The framework of SSLCA algorithm.

incomplete data. Given a model of data generation, and data with some missing values, EM will locally maximize the likelihood of the parameters and give estimates for the missing values. The naïve Bayes generative model allows for the application of EM for parameter estimation. In our scenario, the class labels of the unlabeled data are treated as the missing values.In implementation, EM is an iterative two-step process. Initial parameter estimates are set using standard naïve byes from the labeled instances. The E-step calculate the posterior probability of our parameters the $c_i$ 's, given the $x_i$ and using the current setting of our parameters. Using Bayes rule, we obtain:

$$p(c_i = j|x_i; \Phi, \mu, \Sigma) = \frac{p(x_i|c_i = j; \mu, \Sigma)p(c_i = j; \Phi)}{\sum_{l=1}^{k} p(x_i|c_i = l; \mu, \Sigma)p(c_i = l; \Phi)}$$

Here, $p(x_i|c_i = j; \mu, \Sigma)$ is given by evaluating the density of a Gaussian with mean $\mu_j$ and covariance $\sum_j$ at instance $x_i$, $p(c_i = j; \Phi)$ is given by $\Phi_j$, then the high confidence instances were added into the separate labeled data pool.The M-step estimate new classifier parameters using all the labeled data in the pool. Let $w_i^{(j)} := p(c_i = j|x_i; \Phi, \mu, \Sigma)$ in each iteration, the ratio of each class $\Phi_j$, mean $\mu_j$ and covariance $\sum_j$ were calculated as follow:

$$\Phi_j = \frac{1}{m}\sum_{i=1}^{m} w_i^{(j)}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_i^{(j)} x_i}{\sum_{i=1}^{m} w_i^{(j)}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} w_i^{(j)} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{m} w_i^{(j)}}$$

### 3.2. Entropy split methods

Entropy split is a simple heuristic method to split single views into two views (Du et al., 2011); firstly, is calculate the entropy of each feature in the single view based on the whole data set, which is similar to the entropy calculation when deciding which feature should be chosen as the root of the decision tree.Let $p_i$ denotes the probability of instance belong to class $c_i$ in data set $D$, then $p_i$ can defined as $p_i = |c_{i,D}|/|D|$, so then the expectation information of classification in data set $D$ can be calculated as:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Let $|D_j|/|D|$ denotes the weight of the discrete value $j$ 's division, attribute $A$ has $v$ different value $\{a_1, a_2, \cdots, a_v\}$, so then the entropy of attribute $A$ can be calculated as:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Intuitively, the larger the entropy, the more predictive the feature would be. We simply assign features with the first, third, and so, highest entropy to the first view, then assign features with the second, fourth, and so on, highest entropy to the second view. The rationale is to distribute the high-entropy features evenly in the two views, and thus, both views are more likely to be sufficient.

### 3.3. Choose informative instances

Active learning usually selected uncertainty sample for querying, the active learner queries the instances about which it is least certain how to label. This approach is often straightforward for probabilistic learning models. For example, when using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5.

In our algorithm, we define contribution degree as the selection criteria of informative instances, which not only considering the most uncertain of instances, but also considering the uncertainty difference between the instance and its nearest neighbor, so then the contribution degree can be defined as:

$$Contribution(Conf, x_i) = \frac{1}{N \times Conf(x_i, c)} + \alpha \left| \sum_{x \in N(x_i)} [Conf(x_i, c) - Conf(x, c)] \right|$$

Where, $N$ denotes the number of classification, $Conf(x_i, c)$ denotes the confidence of instance $x_i$ belong to class $c$, $\alpha$ denotes the weight of contribution degree, $N(x_i)$ denotes the nearest neighbor of instance $x_i$. It is based on the idea, if the instance's confidence is high, but the confidence of its neighbor is low, which does not meet the clustering hypothesis; otherwise, those have higher possibility to be informative instances because of high contribution degree.

### 3.4. SSLCA algorithm

SSLCA algorithm is a semi-supervised, multi-view algorithm; the framework is described as Fig. 1. In this section we provide a detailed description, which is shown in Fig. 2.

## 4. Experiment

### 4.1. Experimental datasets

To evaluate the performance of SSLCA algorithm, we employ six UCI data sets (Witten and Frank, 2005) and Chinese Coreference data of natural language processing task in real world.

```
Given:
  -a learning algorithm $\ell$
  -the labeled data $L$ and unlabeled data $U$
  -the number $k$ of iterations to be performed
SSLCA algorithm:
  -split labeled data $L$ and unlabeled data $U$ respectively into view $V_1$ and view $V_2$
  Loop for $k$ iterations
    -use $\ell$ , $V_1(L)$ and $V_2(L)$ to create classifiers $h_1$ and $h_2$
    For each class $c_i$, do
      -let $E_1$ and $E_2$ be the $e$ unlabeled instances on each classifiers $h_1$ and $h_2$
      -selected the high confidence and nearest neighbor instances $E_{hn}$ for $c_i$ , label them according to $h_1$ and $h_2$ ,respectively, and add them to $L$
      -selected the informative instances $E_{in}$ , and add them to data pool $P$
      -remove $E_{hn}$ and $E_{in}$ form $U$
    End For each class $c_i$
  End Loop for $k$ iterations
  -label the instances of data pool $P$ , and add them to $L$
  - create the classifier based on labeled data $L$
```

**Fig. 2.** The SSLCA algorithm.

- UCI data sets: UCI repository currently maintains 239 data sets as a service to the machine learning community. We choose six data sets which all have a common with two classes: breast-w, credit-a, diabetes, ionosphere, kr-vs-kp, sonar. For each data set, we kept the label and put them into the Labeled data pool *L*, remove the label and put them into the unlabeled data pool *U*, the ratio of labeled data and unlabeled data can be set in the next experiment.
- Chinese Coreference data sets: we use the Lancaster Corpus of Mandarin Chinese as the Coreference data, then construct the feature and label the conference relationship by manual operation. In feature extraction, we extract 14 features according the semantic, which contain the single or plural feature, the gender feature, the sentence structure feature, et al. Another, we extract the word and Part-of-Speech feature within five windows of antecedent and anaphor (Zhang, Guo, Yu, Zhang, & Yao, 2009). In the next experiment, we also split those two kinds of feature into two views, and mix these two views together to construct a single-view data set.

### *4.2. Experiment setup*

The performance of SSLCA algorithm is compared with two semi-supervised learning algorithm. The first comparing algorithm is the standard co-training (Blum and Mitchell, 1998). Furthermore, SSLCA is compared with another well-known semi-supervised learning algorithm TSVM (Joachims, 1999). Unlike standard co-training initially trains a classifier on labeled data and then iteratively augment its labeled training set by adding several newly labeled unlabeled instances with most confident predictions of its own. TSVM semi-supervised algorithm initially trains a classifier on labeled data and unlabeled data, which exploit the cluster structure of data and treat it as prior knowledge about the learning task. The SSLCA algorithm uses co-training initially trains a classifier on fever labeled data, which not only utilizes co-training to select the high confidence instances for boosting the classifier, but also exploit informative instances for human annotator.

For any comparing algorithm, several kinds of learning algorithm are employed to perform classifier induction. Specifically, the Naïve Bayes algorithm, the SMO algorithm, and the TSVM algorithm are utilized, which the former two algorithms are come from weka platform, and the svmlight was used as TSVM algorithm, which comes from Martin Theobald's work (Theobald, 2013). Furthermore, the three algorithms are used as baseline algorithms for reference purpose. Naïve Bayes and SMO algorithms train classifiers on only the initial labeled training instances while TSVM trains classifiers on labeled instances together with unlabeled for clustering assumption. For co-training, active learning and SSLCA algorithm equipped with any classifier inducer, 100 independent runs are performed under every configuration of $\ell$. Co-training and active learning algorithms are presented in detail in Section2; SSLCA algorithm is presented in detail in Section 3.

### *4.3. Experiment results on UCI datasets*

We now present our experimental results. The six data sets we used are from the UCI Machine Learning Repository. For all six data sets we choose to increase the amount of labeled data provided, using the rest as unlabeled data. Figs. 3 and 4 illustrates how each comparing algorithm performs with different confidence estimation methods, with the number of labeled training instances increases and different base classifiers.

In Fig. 3, Naïve Bayes is utilized as the confidence estimation method; also it is used as the base classifier of co-training, active learning and SSLCA algorithm. On breast-w dataset, TSVM get the best performance with the increase of labeled instances. In detail, on credit-a dataset, the SSLCA algorithm gets better result when the ratio of labeled data is more than 14%. On diabetes, ionosphere and sonar data sets, the SSLCA algorithm is superior to the other algorithms with the ratio of labeled data increase to certain level. On kr-vs-kp dataset, the SMO algorithm and TSVM algorithm get the best perform when labeled data is small and the SSLCA algorithm get the best performance as the increase of labeled dataset.

In Fig. 4, Expectation Maximization is utilized as the confidence estimation method, and SMO algorithm is used as the base classifier of co-training, active learning and SSLCA algorithm. On breast-w dataset, TSVM get the best performance with the increase of labeled instances. On diabetes, ionosphere and sonar data sets, the SSLCA algorithm get significantly better performance than the other algorithms with the ratio of labeled data increase to certain level. On credit-a and kr-vs-kp datasets, the SMO algorithm and TSVM algorithm get the best perform when labeled data is small and the SSLCA algorithm get the best performance with the increase of labeled dataset.

As shown in Figs. 3 and 4, it is clear that the SSLAC algorithm is superior the traditional supervised algorithm on randomly selected datasets, it also superior the traditional co-training algorithm and active learning algorithm on four datasets, on the other two datasets, it get nearly same performance at least. In summary, the SSLCA algorithm is statistically effective according to the performance on six randomly selected datasets. The algorithm applies co-training to select the most reliable instances for boosting the classifier, but also do not discard the uncertain instances, which usually are the informative instances to the classifier. We label those informative instances and add them into labeled data pool *L* for boosting the classifier, which achieves significant improvement comparing to the other comparing algorithms for sacrificing the same amount of human effort.

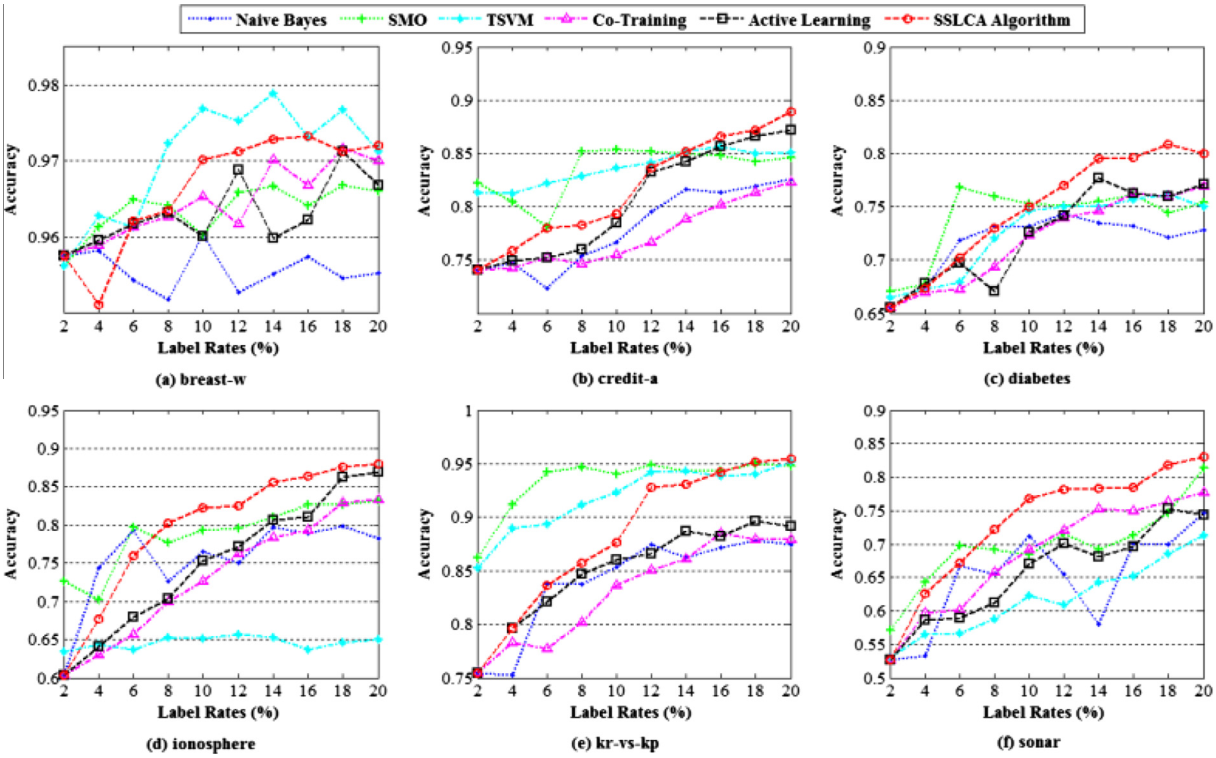

**Fig. 3.** Classification accuracy of each comparing algorithm changes as the number of labeled training instances increases, where Naïve Bayes is utilized as the confidence estimation method.

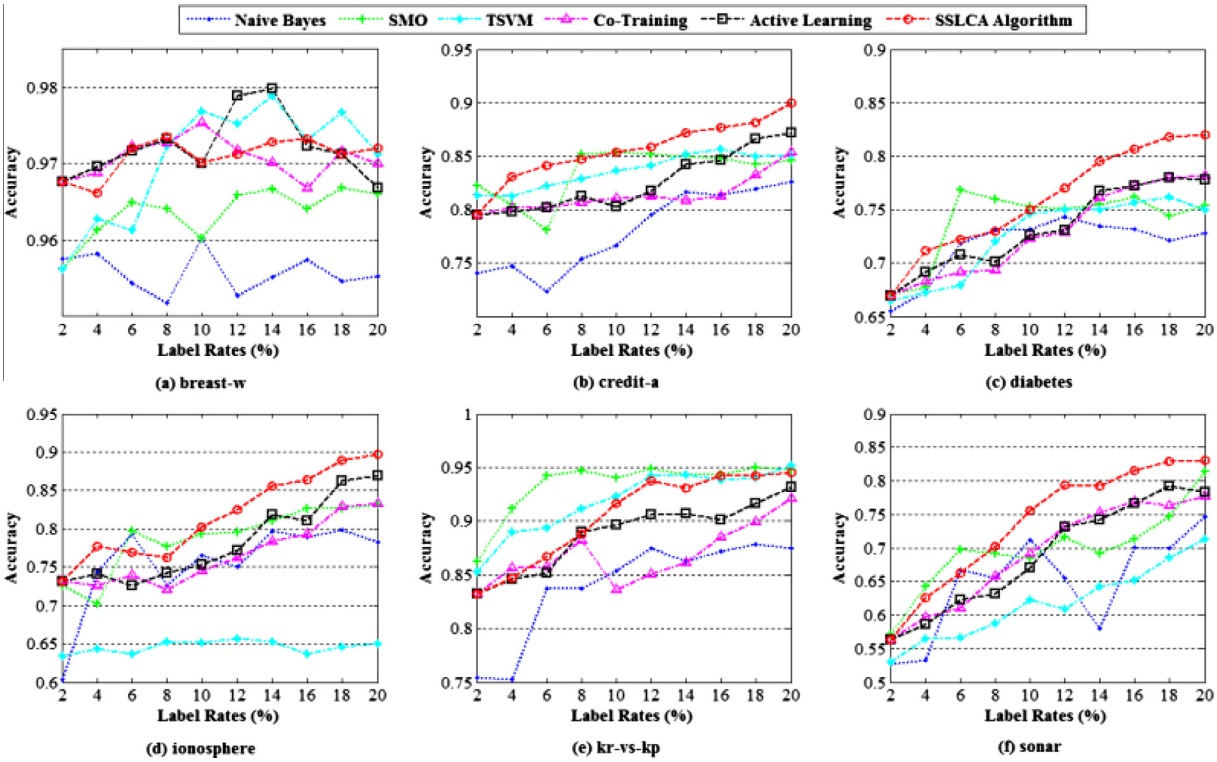

**Fig. 4.** Classification accuracy of each comparing algorithm changes as the number of labeled training instances increases, where Expectation Maximization is utilized as the confidence estimation method.

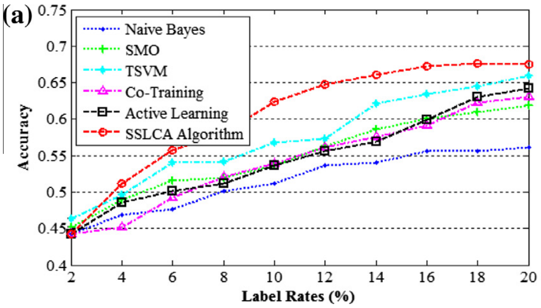

Fig. 5. Classification accuracy of each comparing algorithm changes as the number of labeled training instances increases, where entropy split method is utilized as the view split.

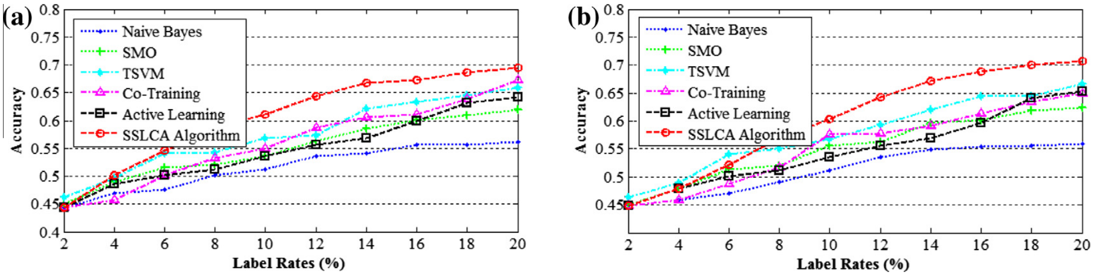

Fig. 6. Classification accuracy of each comparing algorithm changes as the number of labeled training instances increases, where the naturally split views are utilized as the two views.

### 4.4. Application to Chinese Coreference

Chinese Coreference dataset is come from the Lancaster Corpus of Mandarin Chinese. We extract 14 semantic features and word and Part-of-Speech feature within five windows, which constitute the two views of experiment data respectively in co-training and SSLCA algorithm, we also mix these two views together to construct a single-view data set for the other comparing algorithms.

Figs. 5 and 6 illustrates how each comparing algorithm performs with different confidence estimation methods, as the number of labeled training instances increases and different base classifiers. In Fig. 5, entropy split method is utilized as the view split method, (a)Naïve Bayes is utilized as the confidence estimation method and base classifier, (b)Expectation Maximization is utilized as the confidence estimation method and SMO is used as the base classifier of co-training, active learning and SSLCA algorithm. In Fig. 6, the 14 semantic features and word and Part-of-Speech feature are split two views naturally, the experimental configuration of plot (a) and plot (b) is the same as that used in Fig. 5.

As shown in Fig. 5, the SSLAC algorithm is no different than the comparing algorithms when labeled data is small, especially, the TSVM algorithm is superior all comparing algorithms when the ratio of labeled less than 6% in plot (b), but the SSLAC algorithm is consistently superior all comparing algorithms when he ratio of labeled more than 8%. Actually, through observing Fig. 6 we can get the same conclusion that the SSLAC algorithm is consistently superior all comparing algorithms with the increase of labeled data when either Naïve Bayes or SMO is incorporated as the base classifier.

In addition to the above discussion, we further compare the result of Figs. 5 and 6, which can be found that the SSLCA algorithm gets even better result in Fig. 6. This suggests that view split method used in co-training algorithm may play an important role. Moreover, the naturally split views according to semantic features and word and Part-of-Speech feature is more reliable than entropy split method, that is, when there exist sufficient and redundant views, appropriately utilizing them will benefit the learning performance in co-training and the SSLCA algorithm.

## 5. Conclusion

Comparing to the traditional co-training requires the data set can be splits two views, and satisfy the sufficiency and independence assumptions. In practice, those conditions are not easy to achieve. Also co-training usually selects high confidence instances that are labeled with the estimated class labels and add them to the training sets, which does not ensured these selected high confidence instances are more valuable for improving the predictive accuracy. In this paper, we presented a new semi-supervised learning strategy for exploiting the unlabeled data, which combines co-training and active learning. Experiment result demonstrates that our algorithm achieves significant improvement sacrificing the same amount of human effort.

Comparing to the traditional co-training algorithm, our method not only applies co-training to select the most reliable confidence instances for boosting the classifier, but also don't discard the uncertain instances, which usually are the informative instances to the classifier. Comparing to the traditional active learning algorithm, we define contribution degree as the selection criteria of informative instances, which not only considering the most uncertain of instances, but also considering the uncertainty difference between the instance and its nearest neighbor. Moreover, our algorithm is effectively helpful to exploit the informative instances, which can be verified by the fact that the SSLCA algorithm consistently superior the active learning with different base classifier. Compare with co-training is called for precisely when the labeled training set is small, our algorithm worked well on small labeled training sets.

In the future, it is very important to conduct more insightful theoretical analyses on the effectiveness of our approach. Furthermore, designing more appropriate approach as the selection criteria of informative instances is worth studying. The other, effective feature splitting approach is also should make efforts to research and discuss.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to significantly improve the quality of this paper. The research reported in this paper has been partially supported by the National Science Foundations of China under Grant No. 61379158, the Ph.D. Programs Foundation of Ministry of Education of China No. 20120191110028, the National Key Basic Research Program of China (973) No. 2013CB328903.

## References

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training [C]. In *Proceeding of the 11th annual conference on computational learning theory* (pp. 92–100).

Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Du, J., Ling, X., & Zhou, Z. H. (2011). When does co-training work in real data [J]. *IEEE Transactions on Knowledge and Data Engineering, 23*(5), 788–799.

Goldman, S, & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data [C]. In *Proceeding of the 17th international conference on, machine learning* (pp. 327–334).

Joachims, T. (1999). Transductive inference for text classification support vector machines [C]. In *Proceeding of the 6th international conference on, machine learning* (pp. 200–209).

Li, H., Shi, Y., & Liu, Y. (2012). Cross-domain video concept detection: A joint discriminative and generative active learning approach [J]. *Expert Systems with Applications, 39*, 12220–12228.

Lughofer, E. (2012). Hybrid active learning for reducing the annotation effort of operators in classification systems [J]. *Pattern Recognition, 45*(2), 884–896.

Mihalcea, R. (2004). Co-training and self-training for Word Sense Disambiguation [C] In *Proceeding of CONLL-04*.

Settles, B. (2010). Active learning literature survey. Computer sciences Technical Report 1648, University of Wisconsin at Madison, January 26, 2010.

Sun, A., Liu, Y., & Lim, E. P. (2011). Web classification of conceptual entities using co-training [J]. *Expert Systems with Applications, 38*(12), 14367–14375.

Tang, F., Brennan, S., Zhao, Q., Tao, H., Cruz U S., & Cruz, S. (2007). Co-tracking using semi-supervised support vector machines [C]. In *Proceeding of ieee the 11th international conference on computer vision* (pp. 1–8).

Theobald, M. (2013). The program of the svmlight algorithm. http://www.mpi-inf.mpg.de/~mtb/svmlight/JNI_SVM-light-6.01.zip.

Wang, W., & Zhou, Z. H. (2010). A new analysis of co-training [C]. In *Proceeding of the 27th international conference on machine learning, Haifa, Israel*.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann. http://prdownloads.sourceforge.net/weka/datasets-UCI.jar..

Yang, L., Jin, R., & Sukthankar, R. (2009). Bayesian active distance metric learning [C]. In *Proceeding of the 23th conference on uncertainty in, artificial intelligence* (pp. 442–449).

Yu, S., Krishnapuram, B., Rosales, R., & Rao, R. (2011). Bayesian co-training [J]. *Journal of Machine Learning Research, 12*, 2649–2680.

Zhang, Y. H., Guo, J. Y., Yu, Z. T., Zhang, Z. K., & Yao, X. M. (2009). The research on chinese coreference resolution based on maximum entropy model and rules [J], In *Lecture notes in computer science. Vol. 5854* (pp. 1–8).

Zhang, M. L., & Zhou, Z. H. (2011). COTRADE: Confident co-training with data editing [J]. *IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, 41*(6), 1612–1626.

Zhou, Z. H., & Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers [J]. *IEEE Transactions on Knowledge and Data Engineering, 17*(11), 1529–1541.

Zhou, Z. H., & Li, M. (2010). Semi-supervised learning by disagreement [J]. *Knowledge and Information Systems, 24*(3), 415–439.

Zhu, X. J. (2008). Semi-supervised learning literature survey, Computer Sciences TR 1530, University of Wisconsin at Madison, July 19.

Zhu, X. J., Lafferty, J., & Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions [C] In *Proceeding of the ICML-2003 workshop on the continuum from labeled to unlabeled data*, Washington DC.