Evaluation Criteria of Feature Splits for Co-Training

Masahiro Terabe and Kazuo Hashimoto

***Abstract*—Co-training is one of the major semi-supervised learning methods which can induce classifiers from only small number of labeled data and sufficient number of unlabeled data. To perform successfully, co-training requires splitting the features in original data and constructing views. However, it is not so easy to construct good views which satisfy the following conditions for successful performance of co-training: (1) each of views includes good attributes for classification, and (2) the views are sufficiently independent each other. Additionally, some of the recent researches pointed out that co-training can not perform successfully even if the views satisfy the requirements for good performance. We investigate the relationship among the some evaluation criterion for feature splits and the performance of co-training through the well-designed experiments. In this paper, the experimental results and future works for designing ideal artificial feature splits are reported.**

***Index Terms*—co-training, feature splits, dependency, semi-supervised learning.**

I. Introduction

Co-training is one of the major semi-supervised learning methods which induce classifiers from only small number of labeled data and sufficient large number of unlabeled data [2], [5], [20]. Because co-training does not require human efforts to classify the data, the method is applied to various classification learning tasks such as Web page categorization and spam mail filtering.

One of the requirements that co-training performs successfully is as follows: (1) each feature split includes good attributes for classification, and (2) the feature splits are sufficiently independent.

For example, Blum and Mitchell [2] used words in the body of the pages and the words in hyperlinks of other documents referring to that particular page for as natural feature splits. Co-training performs successfully with their adopted natural feature splits. However, in many case, it is not so easy to find such a good natural feature splits which satisfy the requirements for successful performance of co-training.

Some of the researches are conducted to design ideal artificial feature splits. However, the method or criteria to design the ideal artificial feature splits has not been proposed. Additionally, the relationship between the performance of co-training and the characteristics of feature splits has not been investigated sufficiently. Feger et al. [8] reported that co-training shows good performance even if the feature splits are not independent enough.

Final goal of our research is to develop the method or criteria to design ideal artificial feature splits for co-training. As a basic study to achieve the goal, we investigate the relation between the characteristics of feature splits and performance of co-training through well-designed experiments. The characteristics of feature splits are evaluated with some measures. In this paper, the experimental results and feature works for designing ideal artificial feature splits are reported.

Manuscript received January 7, 2008. This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grand-inAid for Scientific Research (B) 19360181, 2007.

M. Terabe is with the Graduate School of Information Sciences, Tohoku University, 6-6-11 Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579, Japan (corresponding author to provide phone: +81-22-795-5856; fax: +81-22-795-5856; e-mail: terabe@ aiet.ecei.tohoku.ac.jp).

K. Hashimoto, is with the Graduate School of Information Sciences, Tohoku University, 6-6-11 Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579, Japan (e-mail: kh@ aiet.ecei.tohoku.ac.jp).

II. Co-Training and Views

A. Co-training Algorithm

The algorithm of co-training [2] is shown in Table 1. Co-training is a semi-supervised learning method which can induce high accuracy classifiers from small labeled and large unlabeled data set.

First, the features in the original data set are split into two feature sets V_1 and V_2, which are called views. As mentioned above, the views should be satisfied the requirements for successful performance of co-training. Two classifiers C_1 and C_2 are learned with using one view of the labeled set L. Here, the two classifiers C_1 and C_2 are slightly different because they are induced with different views though the classifiers are learned from the same data set L and same learner.

Next, all of the examples in unlabeled examples pool U' are classified by the classifier C_1 and C_2 respectively. The 2p+2n examples, which are consisted of p positive and n negative examples with most confidently by C_1 and C_2, are add with predicted label to the labeled data set L. The U' are replenish with examples randomly selected from unlabeled data set U.

By iterating such process, the number of labeled examples L is increased by the self-labeled examples. As a result, the prediction accuracy of the classifiers is improved.

In the framework of co-training, the class of test examples is

predicted by the two classifiers. The confidence of the predicted class is calculated by multiplying the probabilities output by C_1 and C_2. The class with the highest probability is the predicted class by the co-training classifiers.

TABLE 1
CO-TRAINING ALGORITHM

Inputs:

- **L**: a small set of labeled example
- **U**: a small set of unlabeled example
- $\mathbf{V_1}$, $\mathbf{V_2}$: two sets of describing the example

Algorithm:

Create a pool **U'** by randomly choosing **u** examples from **U**

Loop ***k*** iterations
Learn Classifier $\mathbf{C_1}$ from **L** based on $\mathbf{V_1}$
Learn Classifier $\mathbf{C_2}$ from **L** based on $\mathbf{V_2}$

$\mathbf{C_1}$ predicts the class of example from **U'** based on $\mathbf{V_1}$ and chooses the most confidently predicted **p** positive and **n** negative examples $\mathbf{E_1}$
$\mathbf{C_2}$ predicts the class of example from **U'** based on $\mathbf{V_2}$ and chooses the most confidently predicted **p** positive and **n** negative examples $\mathbf{E_2}$
$\mathbf{E_1}$ and $\mathbf{E_2}$ are removed from and added with their labels to **L**
Randomly chose **2p+2n** examples from **U** to replenish **U'**
End

Outputs:

Two Classifiers $\mathbf{C_1}$ and $\mathbf{C_2}$. The confidence of the predicted class is calculated by multiplying the probabilities outputted by $\mathbf{C_1}$ and $\mathbf{C_2}$.

B. Feature Splits and Views

Feature splits and the views are key factor which determines the performance of the co-training.

Blum and Mitchell [2] stated the two requirements for successful co-training. First requirement is that the two views should be sufficiently strong. In other words, the learner can learn the high prediction accuracy classifier using each view individually. Second one is that the two views should be conditionally independent given the class. They proved the requirements theoretically and experimentally.

Nigam and Ghani [13] investigated the effect of dependence between the views to the performance co–training. They found that the co-training performs better on truly independent views than random views through their experiment. This effect was also confirmed by Chan et al. [4].

On the other hands, Feger et al. [8] reported the prediction accuracy of classifiers by co-training does not become better even if the views are truly independent. They concluded that the relation between the characteristics of views and the performance of co-training should be investigated more in detail.

Though many researches are conducted on characteristics of feature splits and the performance of co-training, the method or criteria of designing optimal artificial feature splits has not been proposed. The former researches suggest that there is another feature splits' requirement from ones Blum and Mitchell stated. Thus, we investigated on the requirements through the experiments.

III. EVALUATION MEASURE OF FEATURE SPLITS

The characteristics of feature splits are evaluated with the following measurements.

A. Measure for dependence between two attributes

To evaluate the class-conditional dependence of two views, Feger et al. adopted the Mutual Information (MI) , and we followed them.

The mutual information MI(X,Y) of two features X and Y is depends on their entropy H and defined as follows

$$\begin{aligned} &MI(X,Y) \\ &= H(X) - H(X \mid Y) \\ &= \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{1}{p(x \mid y)} \\ &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Here, H(X) is entropy of X and MI(X,Y) is mutual information between the feature X and Y. If the two features are truly independent, MI(X,Y)=0.

The class-conditional mutual information, CondMI(X,Y) is defined as follows [7].

$$\begin{aligned} &CondMI(X,Y \mid z = c) \\ &= H(X \mid z = c) - H(X \mid Y, z = c) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x,y \mid z = c) \log_2 \frac{p(x,y \mid z = c)}{p(x \mid z = c)p(y \mid z = c)} \end{aligned}$$

Here z=c means that the class z is c. If the feature splits X and Y are truly independent each other, the class-conditional mutual information, CondMI(X,Y) becomes 0.

B. Measure for dependence of two views

As mentioned above, the dependence between the features X and Y can be evaluated by the class-conditional mutual information CondMI(X,Y). Based on class-conditional mutual information, the dependence of inter-views V_1 and V_2 and intra-view V_i is evaluated as follows.

1) Dependence inter-views

The dependence of inter-views V_1 and V_2 is defined by the summation of class-conditional mutual information of the features A_1 and A_2 which belong to the views V_1 and V_2 respectively.

$$InterCMI(V_1, V_2) = \sum_{A_1 \in V_1} \sum_{A_2 \in V_2} CondMI(A_1, A_2)$$

The InterCMI is "lost" mutual information when the feature splits are constructed from the features in the data set. The smaller value of InterCMI is, the dependence inter-views are also smaller.

2) Dependence intra- view

The dependence intra-view V_i is defined with the summation of class-conditional mutual information of the features X and Y which belong to the same view V_i as follows.

$$IntraCMI(V_i) = \sum_{X,Y \in V_i} CondMI(X, Y)$$

This IntraCMI is sum of the mutual information between the features in the same view. The larger value of IntraCMI is, the dependence intra-view is also larger.

C. Measure of strength of feature splits

One of the requirements for views, strength of feature splits means that how the learner can learn the high prediction accuracy classifier using the view individually. We measure the strength of feature splits by the prediction accuracy of the classifier induced using the view from the all of examples.

IV. Experiments

A. Experimental Setup

1) Learning Algorithms

To investigate the relation between the feature splits and the performance of co-training, we conducted the following experiments.

As learners, we use naïve-Bayes (NB) [11], Radial Based Function Neural Networks (RBF NN) [12], [15]. They are also used by Feger et al. [8]. These learners often adopted as basic learner of co-training. They are also used as one of the major application, text classification [3], [4].

In this experiment, we use the learning algorithms implemented in WEKA [17]. All of the parameters of learning algorithms are default ones. The learning parameters of co-Training are also default ones which are used in Blum et al. [2].

2) Data set

The Votes data set is used from UCI Machine Learning Repository [1]. The features of data set are summarized in Table 2. Votes data set has two classes, and all 16 attributes are categorical.

TABLE 2
FEATURES OF VOTES DATASET

- Name: Votes (Congressional Voting Records) data sets.
- # of Instances: 435
- # of Attributes: 16 (All attributes are categorical)
- # of Class: 2

3) Feature splits

We generate a hundred of feature splits randomly. Each feature splits include 8 attributes which are half number of all 16 attributes.

4) The other settings

The number of labeled data given as training data is set to 20. The experiments are done with 10 folded cross validation. The experimental results, such as prediction accuracy of classifiers are average of 10 runs.

B. Results

1) InterCMI and performance of co-training

First, we investigate the relationship between InterCMI and prediction accuracy of classifiers induced by co-training. The smaller InterCMI indicates that the two views are more independent, so it has been expected that the smaller InterCMI feature splits show better performance. We picked up top (smallest) and bottom (largest) 20-InterCMI feature splits respectively, and compare the prediction accuracy of co-trained classifiers.

The experimental results are depicted in Fig. 1. As expected, the prediction accuracy of top 20-feature splits shows better performance than that of bottom 20 in the case of RBF NN learners. In the case of NB learners, the prediction accuracy of top 20 almost the same as that of bottom 20.

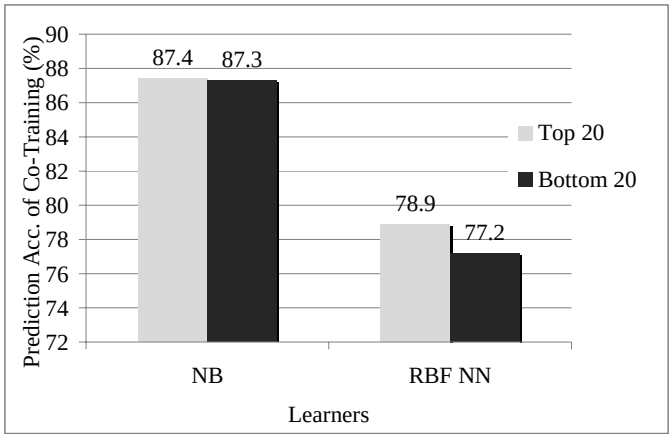

Fig 1. Prediction accuracy of classifiers in top and bottom 20 InterCMI feature splits.

2) Strength of feature splits and performance of co-training

Next, we investigate the effect of strength of feature splits. The strength of feature splits are measured by the prediction accuracy of prediction accuracy of claccifier induced using the view from tha all of examples. Here, the lower prediction accuracy is adopted as measure from two classifiers induced with each views.

We divide the top 20-InterCMI feature splits into two groups based on the strength of feature splits, higher and lower 10 feature splits groups. The experimental results are shown in Fig. 2.

In the case of RBF NN learners, the higher 10 feature strength feature splits group shows better peformance than lower ones. When we adopt the NB as learners, there is not significant difference on the performance between higher and lower groups.

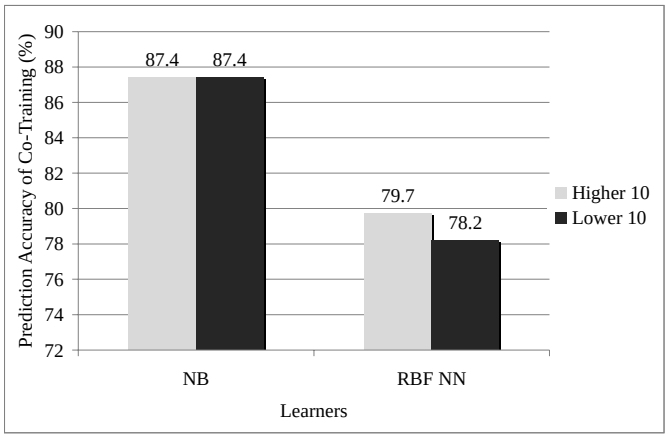

Fig 2. Prediction accuracy of classifiers in higher and lower10 strength of feature splits.

V. Discussion

Blum and Mitchell stated the following two requirements for successful co-training, (1) two views should be sufficiently strong, and (2) two views should be conditionally independent given the class. We confirmed that these two requirements are not valid in some cases.

From our experimental results, we confirmed that the two requirements are valid in the case of RBF NN learners, though the requirements does not hold true in the case of NB learners. These experimental results support the point that co-training does not perform well even though the two views are conditionally independent Feger reported in [8]. The artificial feature splits sometimes does not works well even though the feature splits satisfies the two requirements stated by Blum and Mitchell.

For developing the method or criteria to design ideal artificial feature splits for co-training, more exploration of additional requirements and their measure should be conducted.

VI. Conclusion

In this paper, we investigate the relation between the characteristics of feature splits and performance of co-training as a basic study to develop the method or criteria to design the ideal artificial feature splits.

To confirm the effectiveness of two requirements for successful co-training stated by Blum and Mitchell, we conducted the experiments with many kinds of feature splits, and investigate the relationship between performance of co-training and the characteristic measure of feature splits. The experimental results show that the two requirements stated by Blum et al. are not valid in some cases.

To design the ideal artificial feature splits, some more explorative study of additional requirements and their measure is required. The following issues are our future works.

- **Adopting various kinds of learners:** we conducted the experiments with naïve-Bayes and RBF neural networks as learners, and the performances are different. To clarify the difference of performance among learners and induce general requirements, we should conduct experiments with the other learners such as support vector machines (SVM) [6][16] which are adopted as learners for co-training for text classification problem [4][8].
- **Experiments with real and large data sets:** The data set we use in this paper has only 16 attributes. The real data sets such as text classification problem, always has much more attributes. Additionally, more unlabeled data sets are prepared. The relation ship between performance of co-training and feature splits in the case of large data sets is necessary to investigate for applying to the real problem.
- **Exploring another measure of feature splits:** In this paper, we adopted the measurements of feature splits proposed by Fager et al. [8]. Exploration of effective measurement which represents the characteristics of feature splits is required, especially for the numerical attributes.

References

[1] A. Asuncion and D. J. Newman. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science, 2007.

[2] A. Blum and T. Mitchell, “Combining Labeled and Unlabeled Data with Co-Training”, In Proceedings of the eleventh annual conference on Computational Learning Theory (COLT’98), pp. 92-100, 1998.

[3] A. Bouchachia, “RBF Networks for Learning from Partially Labeled Data”, in Proceedings of the 22nd ICML Workshop on Learning with Partially Classified Training Data, pp.10-18, 2005.

[4] J. Chan, I. Koprinska, and J. Poon, “Co-training with a Single Natural Feature Set Applied to Email Classification”, In Proceedings of IEEE/WIC/ACM Conference on Web Intelligence (WI’04), 2004.

[5] O. Chapelle , B. Scholkopf , and A. Zien, Semi-Supervised Learning, MIT Press, 2006.

[6] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition", *IEEE Trans. on Electronic Computers*, Vol. EC-14, No.3, pp.326-334, 1965.
[8] F. Feger and I. Koprinska, "Co-training Using RBF Nets and Different Feature Splits", In Proceedings of 2006 International Joint Conference on Neural Network, pp. 1878-1885, 2006.
[9] S. Kubo, M. Terabe, K. Hashimoto, R. Ono, "Tri-Testing: A novel Semi-Supervised Learning Method based on Ensemble Learning and Active Learning", In Proceedings of the Sixth Mexican International Conference on Artificial Intelligence (MICAI'07), 2007.
[10] M. Li and Zhi-Hua Zhou, "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers", IEEE Trans. on Knowledge and Data Engineering, Vol.17, No.11, Nov. 2005, pp.1529-1541, 2005.
[11] T. Mitchell, Machine Learning, McGraw-Hill, 1997.
[12] J. Moody and C. Darken, "Fast Learning in Networks of Locally-Tuned Processing Units", Neural Computation 1 pp. 289-303, 1989.
[13] K. Nigam and R. Ghani, "Understanding the Behavior of Co-Training", in Proceedings of the Workshop on Text Mining at the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000), 2000.
[14] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training", In Proceedings of the Ninth International Conference on Information and Knowledge (CIKM'00), pp. 86–93, 2000.
[15] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks", *Neural Computation*, Vol.3, No. 2 (Mar. 1991), pp. 246-257, 1991.
[16] V. Vapnik, Statistical Learning Theory, Wiley, 1998.
[17] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques (2nd Edition), Morgan Kaufmann, San Francisco, 2005.
[18] Y. Yang and J. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", In Proceedings of Fourteenth International Conference on Machine Learning, pp.412-420, 1997.
[19] Y. Zhou and S. Goldman, "Democratic Co-Learning", In Proceedings of the Sixteenth IEEE international Conference on Tools with Artificial intelligence (ICTAI'04), pp. 594-202 2004.
[20] X. Zhu, "Semi-Supervised Learning Survey", Computer Sciences TR 1530, University of Wisconsin Madison, 2007. Available at http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html.