Analysis of Co-training Algorithm with Very Small Training Sets

Luca Didaci, Giorgio Fumera, and Fabio Roli

Department of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi 09123, Cagliari, Italy
{luca.didaci,fumera,roli}@diee.unica.it
http://prag.diee.unica.it/

Abstract. Co-training is a well known semi-supervised learning algorithm, in which two classifiers are trained on two different views (feature sets): the initially small training set is iteratively updated with unlabelled samples classified with high confidence by one of the two classifiers. In this paper we address an issue that has been overlooked so far in the literature, namely, how co-training performance is affected by the size of the initial training set, as it decreases to the minimum value below which a given learning algorithm can not be applied anymore. In this paper we address this issue empirically, testing the algorithm on 24 real datasets artificially splitted in two views, using two different base classifiers. Our results show that a very small training set, even made up of one only labelled sample per class, does not adversely affect co-training performance.

Keywords: Semi-supervised learning, Co-training, Small sample size.

1 Introduction

Semi-supervised learning (SSL) methods are useful in many practical applications in which a small set of labelled samples L is available, but a large set of unlabelled samples U can be exploited to improve the performance of learning algorithms. Typical examples are text (e.g., Web page) classification, and biometric authentication. Co-training is a well known SSL algorithm originally proposed in [1], for binary classification problems in which two different views (feature spaces) X^1 and X^2 are available. Starting from a small training set $L = \{(x_i^1, x_i^2, y_i)\}_{i=1}^{n_L}$, where $x_i^1 \in X^1$, $x_i^2 \in X^2$ and $y \in \{-1, +1\}$, it consists of iteratively re-training a pair of classifiers $f^1 : X^1 \to Y$ and $f^2 : X^2 \to Y$, adding to L at each step the unlabelled samples from a given set $U = \{(x_i^1, x_i^2)\}_{i=1}^{n_U}$ that are classified with high confidence by one of the classifiers. Under the assumption of conditional independence between the views, given the class Y (i.e., $P(X^1, X^2|Y) = P(X^1|Y)P(X^2|Y)$, and of sufficiency of each view (i.e., the classes can be perfectly discriminated in each view, if there were a sufficient number of samples), it was shown that co-training allows both classifiers to attain the same performance as they were trained on a large set of labelled samples.

Several authors have theoretically or empirically investigated several aspects of co-training; for instance, how it works when the original assumptions do not hold (which

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 719–726, 2012.

can often happen in practice), or how it works under different and less restrictive assumptions. Balcan et al. [2] provide a PAC-style analysis, and showed that it is possible to relax the condition of independence between the views, if the classifiers are never "confident but wrong". Christoudias el al. [3] examined settings in which classifiers are not compatible due to noise in one view (i.e., even in a ideal situation with a very large training set, their decisions disagree). Didaci and Roli [4] investigated, form a Bayesian point of view, the consequence of the non-sufficiency of the views. Du et al. [5] investigated the possibility of predicting whether co-training will work or not for a given problem, and whether single-view problems can be artificially decomposed into two views to exploit co-training. Zhou et al. [6] considered the limit case when only one labelled sample (x, y) is available, and thus co-training can not be applied, and proposed a method for adding to (x, y) artificially labelled samples to enable the application of co-training.

In this work we address a different aspect of co-training behaviour that has been overlooked in the literature so far, namely, how it performs as the size of L decreases toward the minimum possible value, below which a given learning algorithm can not be applied anymore. As an example, to estimate the covariance matrix of a linear or quadratic Gaussian classifier, at least two samples are needed for each class. In this paper we address this issue empirically, using 24 single-view data sets from the UCI Repository [7], and two different base classifiers. Our goal is to provide a first answer to the questions of whether, and to what extent the performance of co-training is affected by the size of the initial training set L.

After a summary of the co-training algorithm and of previous works in Sect. 2, in Sect. 3 and 4 we present the results of our experiments. Conclusions are drawn in Sect. 5.

2 Background

2.1 The Co-training Algorithm

In this paper we consider the standard version of the co-training algorithm given in [1], which is reported as Algorithm 1. First, a subset U' of unlabelled samples is randomly selected from the available data set U. Then, the following steps are repeated for a predefined number of iterations. Two separate classifiers, f^1 and f^2, one for each view, are trained on the initial, small, training set L, and are then used to label the samples in U'. For each classifier, the p samples of class $+1$ and the n samples of class -1 that are labelled with the highest confidence among the ones of U' are added to L. Classifier confidence can be evaluated, for instance, as the estimated posterior probability, while p and n are chosen such that they are proportional to the (estimated) class priors. The selected $2p + 2n$ samples are then removed from U', and other $2p + 2n$ samples are randomly selected from U and added to U'. The reason of using a subset U' of the unlabelled samples U is that this forces f^1 and f^2 to select samples that are more representative of the underlying distribution, even if they may be not the ones labelled with highest confidence among the ones in U [1].

Previous works on co-training mainly considered Naive Bayes and decision trees as base classifiers. Nevertheless, to the purpose of this work, we point out that each base

Algorithm 1. Co-training algorithm

Input: L and U: sets of n_{L} labelled and n_{U} unlabelled samples, respectively, represented in two views X^1 and X^2; k: number of iterations; $n_{U'} < n_U$: number of samples to be drawn from n_{U}; n, p: number of pattern selected by each classifier at each step. n, p are proportional to priors.
Output: two trained classifiers $f^1 : X^1 \rightarrow Y$ and $f^2 : X^2 \rightarrow Y$.

```
$U' \leftarrow$ a set of $n_{U'}$ samples randomly drawn from $U$
for $k$ iterations do
    Train a classifier $f^1$ on the $X^1$ view of $L$, and a classifier $f^2$ on the $X^2$ view of $L$
    for $i = 1, 2$ do
        Let $f^i$ labels all samples in $U'$
        $U'_i \leftarrow$ the $p$ samples labelled as $+1$ and the $n$ ones labelled as $-1$ with higher confidence by $f^i$
        $L \leftarrow L \cup U'_i, \;\; U' \leftarrow U' - U'_i$
    end for
    Randomly choose $2p + 2n$ samples from $U$, and move them to $U'$
end for
```

classifier has an intrinsic limit to the minimum number of labelled samples of each class in the training set, that allows the corresponding learning algorithm to be applied. We denote these values as $|L^+|_{\min}$ and $|L^-|_{\min}$, respectively for $y = +1$ and $y = -1$. Usually $|L^+|_{\min} = |L^-|_{\min}$. In some cases $|L^+|_{\min} = |L^-|_{\min} = 1$ (e.g., a support vector machine), while in other cases both values can be greater than 1 (e.g., Gaussian classifiers).

2.2 Previous Works

Among previous works on co-training, we mention here the ones that are most related to this paper.

In [2] it was shown that the assumption of conditional independence given the class label can be relaxed, provided that the learning algorithm is never "confident but wrong", i.e., it never misclassifies samples with high confidence. This result could in principle be exploited to make co-training work even when the initial L is very small, as discussed in Sect. 4.

In [6], the limit case when only one labelled sample is available was considered, namely $|L| = 1$. This can happen in applications like content-based image retrieval, and online web-page recommendation. In this case the standard co-training algorithm can not be applied, since it requires a binary base classifier. The proposed solution is first to label and add to L some samples of U, using a different SSL method, such that both classes are represented, and then run co-training, starting from the updated L. The resulting performance of co-training was evaluated for some different sizes of L. However, sizes very close, or equal, to the minimum one required to run the considered base classifiers were not considered.

Finally, in [5] the possibility of predicting whether co-training will work for a given problem was investigated. To this aim, methods for estimating whether the original assumptions of [1] hold or not were devised, using the samples of L. The conclusion was that no reliable estimate can be obtained from a small L. The related problem of artificially decomposing single-view problems (with a small set of labelled samples) into two views, to exploit co-training, was addressed in the same work. No effective method was found to this aim. The reason is that this requires to find the "best" split of the original feature set, according to the co-training underlying assumptions. However, the validity of such assumptions can not be determined from a small L.

In our experiments we will exploit the methods of [5] to artificially split the considered data sets into two views, since for our purposes the best split can be estimated using the whole labelled data sets.

3 Experimental Setup

Co-training was implemented as in Algorithm 1, with $|U'| = 0.3|U|$, similarly to [1,5]. We chose the values of p and n, such that p/n is (approximately) equal to the estimated class priors. According to [2], to this end we chose the smallest possible p and n. We set the number k of co-training iterations in order to allow co-training to collect all samples in U. The exact value of k depends thus on the size of the data set, and on the values of n and p.

Two different base classifiers were used: Naive Bayes (NB), and K-nearest neighbors (K-NN), with $K = 1$. In the case of real-valued features, NB was implemented by subdividing their range into 10 bins of equal width. The experiments have been carried out on 24 single-view two-class data sets, previously used in [5]. They have been artificially subdivided into two views using the method proposed in [5], which aims at minimising the correlation between the corresponding feature subsets, given the class, and maximising the separability of classes in each view, to meet as much as possible the assumptions of [1].

Ten different runs of the experiments have been made. At each run, each data set was randomly subdivided into a labelled training set L, an unlabelled data set U, and a testing set. We considered different sizes of L, as explained below. The size of the testing set was 25% of the entire data set, whilst the remaining data was used as the set U.

The goal of our experiments was to analyse the behaviour of co-training, as the size of L decreases to $|L|_{\min} = |L^+|_{\min} + |L^-|_{\min}$. Note that, with the chosen base classifiers, $|L^+|_{\min} = |L^-|_{\min} = 1$. To this aim, we considered values of $|L|$ ranging from 2 (i.e., $|L^+| = |L^-| = 1$) to 50% of the entire data set. L was obtained by stratified sampling from the whole data set, i.e., $|L^+|$ and $|L^-|$ were chosen such that $|L^+|/|L^-|$ was (almost) equal to the original proportion between the two classes. When the size of L was reduced to the extent that the corresponding $|L^+|$ or $|L^-|$ (chosen as explained above) attained its lowest possible value (respectively, $|L^+|_{\min}$ and $|L^-|_{\min}$), then the most populated class was undersampled. Note that, in this case, L was not representative of the underlying class priors.

At each run, and for each given L, we run co-training and computed its testing set performance. We then checked whether co-training performance attained for $|L| = |L|_{\min}$ was better than the performance attained by the corresponding base classifier trained on the same L, without co-training. If not, we checked for which size of L (if any) co-training outperformed the base classifier trained on the same L, without co-training. We considered two performances significantly different, if the difference between their average values over the ten runs was higher than the sum of the corresponding standard deviations, divided by the square root of the number or runs.

4 Experimental Results

In Table 1 we report the characteristics of the data sets used in the experiments. Table 2 shows the comparison between co-training performance attained for $|L| = |L|_{\min}$, and the performance attained by the corresponding base classifier trained on the same L, without co-training, for both classifiers. The meaning of table entries is the following: 0: no statistically significant performance difference in both views; 1: co-training was outperformed by the base classifier in both views; 1*: co-training was outperformed by the base classifier in one view, no statistically significant performance difference in the

Table 1. Characteristics of the data sets used in the experiments. An asterisk after the data set name denotes that its classes have been merged into two artificial classes, as in [5]. The numbers between brackets in the "n. samples" column denote the number of samples per class. The column "views size" reports the number of features in each view, obtained with the method of [5].

ID	Dataset	n.features	n. samples	views size
1	Audiology*	55	200 [48, 152]	31/24
2	Automobile*	24	193 [130, 63]	22/2
3	Breast Cancer W.	8	699 [458, 241]	5/3
4	Winsconsin D.	30	569 [212, 357]	11/19
5	Winsconsin Progn. 1	33	194 [46, 148]	12/22
6	Winsconsin Progn. 2	32	198 [47, 151]	13/19
7	Contraceptive Method	9	1473 [629, 844]	2/7
8	Horse colic	5	368 [232, 136]	4/1
9	Credit Approval	15	653 [296, 357]	9/6
10	Dermatology*	33	366 [112,254]	17/16
11	Pima Indians Diabetes	8	768 [500, 268]	6/2
12	E.Coli*	7	336 [143/193]	3/4
13	Flags*	28	194 [134/60]	15/13
14	Heart (Cleveland)*	11	303 [164, 139]	5/6
15	Heart (LongBeach)*	4	200 [144, 56]	2/2
16	Heart-statlog	13	270 [150, 120]	4/9
17	Hepatitis_1	19	80 [13, 67]	12/7
18	Ionosphere	33	351 [225, 126]	6/27
19	Chess (King Rook vs King Pawn)	36	3196 [1669, 1527]	13/23
20	SolarFlare_2*	10	1389 [1321, 68]	8/2
21	Sonar Mines vs. Rocks	60	208 [97, 111]	34/26
22	Spambase	57	4601 [2788, 1813]	21/36
23	Splice2*	60	3186 [1532, 1654]	32/28
24	Tic-Tac-Toe	8	958 [626, 332]	3/2

Table 2. Comparison between the average co-training performance attained for $|L| = |L|_{\min}$, and the average performance attained by the corresponding base classifier trained on the same L, without co-training (see the text for the meaning of table entries)

ID	Dataset	Naive Bayes	K-NN
1	Audiology*	2	2*
2	Automobile*	2*	2*
3	Breast Cancer W.	2	2
4	Winsconsin D.	2	2
5	Winsconsin Progn. 1	2*	2
6	Winsconsin Progn. 2	2	2
7	Contraceptive Method	1*	1
8	Horse colic	0	1*
9	Credit Approval	2	0
10	Dermatology*	2	2
11	Pima Indians Diabetes	0	2*
12	E.Coli*	2	0
13	Flags*	2	2
14	Heart (Cleveland)*	2	0
15	Heart (LongBeach)*	2*	2
16	Heart-statlog	2	-
17	Hepatitis_1	2	2*
18	Ionosphere	2	0
19	Chess (King Rook vs King Pawn)	1	1*
20	SolarFlare_2*	2	2
21	Sonar Mines vs. Rocks	0	0
22	Spambase	2	0
23	Splice2*	0	1*
24	Tic-Tac-Toe	0	2

other view; 2: co-training outperformed the base classifier in both views; 2*: co-training outperformed the base classifier in one view, no statistically significant performance difference in the other view; -: co-training outperformed the base classifier in one view, and was outperformed by the base classifier in the other view.

When Naive Bayes was used, co-training outperformed in both views the base classifier trained only on L, in 14 data sets. It was instead outperformed in both views only once (Chess data set). The performance was similar in the remaining 5 data sets. When the K-NN classifier was used, results were similar: co-training was better than the base classifier, on both views, in 9 data sets; it was outperformed by the base classifier in one data set (Contraceptive Method); their performance was similar in the other 6 data sets.

We then evaluated co-training performance as above, for $|L| > |L|_{\min}$. The results (not reported here due to lack of space) showed that, when co-training outperformed in both views the base classifier for $|L| = |L|_{\min}$, the same happened for $|L| > |L|_{\min}$. A representative example of this behaviour is reported in Fig. 1 for the Hepatitis_1 data set. Note that co-training performance is almost constant for all the considered $|L|$ values.

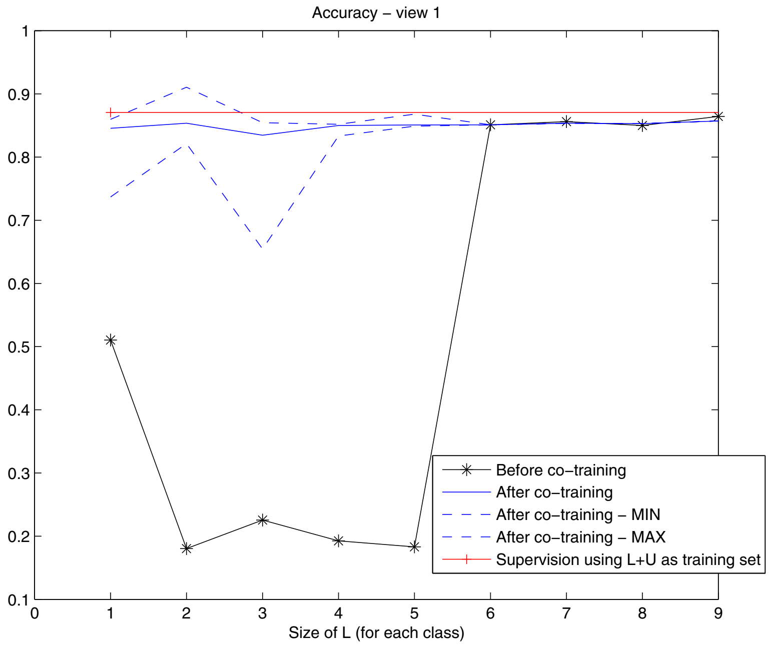

Fig. 1. Co-training performance on the Hepatitis_1 data set, as a function of $|L^+| = |L^-|$, using the NB classifier. "Before co-training": average performance of the base classifier trained on L. "After co-training": average co-training performance, starting from the same L. The minimum (MIN) and (MAX) co-training accuracy is also reported, over the ten runs. For reference, the average performance of the base classifier trained on $L + U$, using the true class labels of the samples in U, is also shown ("Supervision").

These results suggest that co-training performance seems not affected by the size of L, and that co-training can work (i.e., can outperform the base classifier trained only on L) also for very small $|L|$ values, including $|L| = |L|_{\min} = 2$.

5 Conclusions

We addressed the issue of evaluating co-training performance as a function of the size of the labelled training set L, as it decreases to the minimum value below which the considered base classifier can not be applied anymore. Results attained on 24 real data sets, artificially splitted into two views, using two different base classifiers, showed that: (i) co-training performance seems not affected when L reduces to the smallest set of samples that allows the chosen learning algorithm to run; (ii) it can outperform the base classifier trained on L, for any size of L, and, in particular, also in the limit case $|L| = |L|_{\min}$. In other words, co-training can work even with one sample per class. This behaviour could be explained by the results of [2], mentioned in Sect. 2.2. Even if the two base classifiers trained on the initial L have a poor accuracy (which is likely to happen, when L is very small), adding to L *only a few* unlabelled samples that are classified with the highest confidence may prevent from introducing mislabelled samples in the updated training set. This allows the training set to become more representative of the underlying distribution at each iteration, especially if the two views are independent, or at least exhibit a low correlation. Accordingly, a very conservative updating policy of L could be beneficial to co-training, when L is very small.

We point out that these results do not allow one to reliably *predict* whether co-training will work, for a given, real data set, i.e., whether it will outperform the base classifier trained on L. This remains an open issue, as shown in [5]. Our results nevertheless provide evidence that a very small L is not an adverse factor for co-training performance.

Acknowledgment. This work has been partly supported by the project CRP-18293 funded by Regione Autonoma della Sardegna, L.R. 7/2007, Bando 2009.

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings 11th Annual Conference on Computational Learning Theory, pp. 92–100. ACM (1998)
2. Balcan, M.F., Blum, A., Yang, K., Saul, L.K.: Co-Training and Expansion: Towards Bridging Theory and Practice. In: Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 89–96. MIT Press (2005)
3. Christoudias, C.M., Urtasun, R., Kapoorz, A., Darrell, T.: Co-training with noisy perceptual observations. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2844–2851 (2009)
4. Didaci, L., Roli, F.: A Bayesian Analysis of Co-Training Algorithm with Insufficient Views. In: Proc. 11th International Conference on Information Science, Signal Processing and their Applications, pp. 1141–1145. IEEE (2012)
5. Du, J., Ling, C.X., Zhou, Z.-H.: When Does Co-Training Work in Real Data? IEEE Transactions on Knowledge and Data Engineering 23(35), 788–799 (2011)
6. Zhou, Z.-H., Zhan, D.-C., Yang, Q.: Semi-Supervised Learning with Very Few Labeled Training Examples. In: Proc. AAAI, pp. 675–680 (2007)
7. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2010), `http://archive.ics.uci.edu/ml`