Unlabeled Data and Multiple Views

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
zhouzh@lamda.nju.edu.cn

Abstract. In many real-world applications there are usually abundant unlabeled data but the amount of labeled training examples are often limited, since labeling the data requires extensive human effort and expertise. Thus, exploiting unlabeled data to help improve the learning performance has attracted significant attention. Major techniques for this purpose include semi-supervised learning and active learning. These techniques were initially developed for data with a single *view*, that is, a single *feature set*; while recent studies showed that for multi-view data, semi-supervised learning and active learning can amazingly well. This article briefly reviews some recent advances of this thread of research.

1 Introduction

Traditional supervised learning approaches try to learn from *labeled* training examples, i.e., training examples with ground-truth labels given in advance. In many real-world tasks, however, there are often abundant *unlabeled* data but limited amount of labeled training examples. Simply neglecting the unlabeled data would waste useful information, while learning only from the limited labeled data would be difficult to achieve strong generalization performance. Thus, it is natural that exploiting unlabeled data to help improve learning performance, especially when there are just a few training examples, has attracted significant attention during the past decade.

Major techniques for this purpose include *semi-supervised learning* and *active learning*. Semi-supervised learning [5, 30, 28] tries to exploit unlabeled data in addition to labeled data automatically, without human intervention; while active learning [17] assumes interaction with an *oracle*, usually human experts, by trying to minimizing the number of queries on ground-truth labels for constructing a strong learning model. Semi-supervised learning can be divided further into *pure* semi-supervised learning which takes an open-world assumption that the trained model may be applied to unseen unlabeled data, and *transductive learning* which adopts a closed-world assumption that the test instances are exactly the given unlabeled data. The idea of transductive learning can be traced back to [20], where it was argued that we do not need to optimize the learning performance on the whole instance space if we only care the generalization performance on a specific set of test instances.

F. Schwenker and E. Trentin (Eds.): PSL 2011, LNAI 7081, pp. 1–7, 2012.

Data in many tasks have only a single *view*, i.e., a single feature set, and each instance is described by a single feature vector in such situations. However, there are also many real-world tasks where the data have multiple views, i.e., multiple feature sets, and each instance is described by multiple feature vectors in different feature spaces simultaneously. For example, a web page can be classified based on information appearing in the web page itself, or based on anchor texts pointing to this web page; thus, features describing the information in the web page itself constitute the first view, while features describing the information in the anchor texts constitute the second view. Another example is multimedia data, where text features, image features and audio features constitute three different views, respectively. Formally, a single-view example appears as $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x_i}$ is the instance and y_i is the class label; while a multi-view example appears as $([\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}], y_i)$ where $[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}]$ is an instance pair in different views (e.g., $\boldsymbol{x}_{i1}$ is a text feature vector while $\boldsymbol{x}_{i2}$ is an image feature vector). Rather than simply concatenating $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ into a single instance, *multi-view learning* deals with multi-view data by exploiting the views.

Semi-supervised learning and active learning techniques were initially developed for single-view data. It has been found that, however, for multi-view data, semi-supervised learning and active learning can work amazingly well. This article briefly reviews some recent advances of this thread of research.

2 Semi-supervised Learning and Multi-view

Among mainstream semi-supervised learning techniques, the disagreement-based approaches are particularly interesting. These approaches train multiple learners for the task and exploit the disagreements among the learners during the semi-supervised process [28]. A representative is the co-training approach [3] which works with two views. This approach trains a classifier from each view, respectively, using the original labeled data. Then, each classifier selects and labels some highly-confident unlabeled instances to refine its peer classifier. The whole process repeats until no classifier changes or a pre-set number of learning rounds have been executed.

Such a learning process is simple yet effective, and it has many variants and applications [28]. Theoretically, Blum and Mitchell [3] proved that if the two views are "sufficient and redundant" (i.e., each view contains sufficient information for constructing a strong classifier while the two views are conditionally independent given the class label), the predictive accuracy of an initial weak classifier can be boosted to arbitrarily high using unlabeled data by co-training. Dasgupta et al. [7] showed that the generalization error of co-training is upper-bounded by the disagreement between the two classifiers. In real-world tasks, however, the requirement of sufficient and redundant views is too luxury. Actually, even for the motivating example of web page classification task given in [3], it is arguable that whether the requirement holds or not. Thus, researchers tried to find relaxed conditions for co-training to work.

Abney [1] showed that the two views are not needed to be conditionally independent, and a "weak independence" assumption is sufficient for co-training to work. Balcan et al. [2] proved that even the weak independence is not needed if PAC learners can be obtained on each view, and a weaker assumption of "expansion" of the underlying data distribution is sufficient for co-training to work. All the above analyses assumed two views. Wang and Zhou [21] disclosed that for PAC learners, the key for co-training-style approaches is the existence of a "large difference" between the two learners, while it is unimportant whether the difference is achieved by using two views or from other channels. This result provides theoretical support to single-view variants of co-training which work well without two views by training the two learners using different learning algorithms [9], different parameter configurations [27, 10], etc.

As introduced above, more and more relaxed *sufficient conditions* for co-training have been discovered; however, the *sufficient and necessary condition* remained unknown for over ten years. Recently, through establishing a connection between the two mainstream semi-supervised learning approaches, that is, disagreement-based and graph-based approaches, Wang and Zhou [24] addressed this problem. They showed that the co-training process is equivalent to a combinative label propagation process over graphs corresponding to the two views, and thus, sufficient and necessary conditions for co-training were discovered by analyzing the properties of the corresponding graphs under different situations. Wang and Zhou [24] also proved a *necessary condition*, which discloses that the existence of two views is not really needed for co-training-style approaches.

Now it is known that multi-view is neither necessary [24] nor "tightly" sufficient [21] for co-training-style approaches; however, when the data have multiple views, amazing performances can be achieved. For example, Zhou et al. [29] showed that, with sufficient and redundant views, it is possible to execute an effective semi-supervised learning with a single labeled training example, owing to helpful information contained in the correlation between the two views.

3 Active Learning and Multi-view

Active learning generally tries to query the labels of unlabeled *informative* instances (e.g., [18]) or *representative* instances (e.g., [6]). Recently there are some proposals of querying on *informative and representative* unlabeled instances (e.g., [11]). Those principles can be accomplished in different ways, leading to different active learning approaches. A simple yet effective multi-view active learning approach, co-testing [14], trains two classifiers each from one view and then picks their most disagreed unlabeled instance to query, with the intuition that the most disagreed unlabeled instance would be the most informative for improving learning performance.

Theoretically there are two situations of active learning; that is, *realizable* active learning where the data can be perfectly separated by a hypothesis in the

hypothesis class, and *non-realizable* active learning where the data cannot be perfectly separated by any hypothesis in the hypothesis class because of noise. For the realizable case, many studies showed that *exponential* improvement in sample complexity can be achieved by active learning. Wang and Zhou [22] proved that an multi-view active learning approach can also improve the sample complexity remarkably in realizable case.

The realizability assumption, however, rarely holds in real practice, and the non-realizable case is more important since it is closer to real setting. Kääriäinen [12] showed that the lower bound of general non-realizable active learning is in the same order as the upper bound of *passive learning* (i.e., common supervised learning), implying that active learning in the non-realizable case is not as helpful as that in the realizable case if nothing is known about the noise model. In analyses on non-realizable active learning, the Tsybakov noise model [19] becomes more and more popular. It is known that exponential improvement in sample complexity is achievable with *bounded* Tsybakov noise, but for *unbounded* Tsybakov noise which is more closer to real settings, several researchers such as Castro and Nowak [4] concluded that it is hard to achieve exponential improvement, or in other words, active learning would not be remarkably helpful. Recently, Wang and Zhou [23] proved that an multi-view active learning approach can exponentially improve the sample complexity in non-realizable case with unbounded Tsybakov noise. This is a good news, implying that active learning is possible to help remarkably if specific data properties are adequately considered and exploited.

It is not difficult to combine multi-view active learning with semi-supervised learning. For example, Muslea et al. [15] combined co-testing with co-EM [16], a probabilistic variant of co-training, where co-EM iteratively learned two classifiers each from one view by exploiting unlabeled data, and the unlabeled instances disagreed by the two classifiers were selected to query by co-testing. Empirical studies showed that such an approach performed better than semi-supervised learning. Zhou et al. [25] proposed a single-view active semi-supervised learning approach for content-based image retrieval. They generated two learners from labeled images using different parameter configurations. Each learner attempts to assign a rank to unlabeled images in the imagebase, and then passes some irrelevant images with high confidence to its peer as additional negative examples. The two learners are updated and such a process repeats. At the meanwhile, rather than passively waiting for user feedback, a pool of images is actively prepared for the user to give feedback. The pool is composed of images on which the two learners are both with high confidences but disagree, or both with low confidences. The whole process leads to the *active semi-supervised relevance feedback* scheme which is useful for information retrieval tasks. Theoretically, Wang and Zhou [22] proved that an multi-view active semi-supervised learning approach is able to exponentially improve sample complexity in contrast to pure semi-supervised learning.

4 About the Views

Different assumptions can be made for the views, from the possibly weakest that "each view contains information for training weak classifiers that are slightly better than random guess", to the possibly strongest that the views are "sufficient and redundant".

View split, i.e., splitting a single view into multiple views, is a possible solution for applying multi-view approaches to single-view data. It was shown in [16] that for data with a lot of redundant features, such as text data, a random split of the features is able to generate two views to enable standard co-training. It is evident, however, that a random split would not work in most cases. Du et al. [8] tried several heuristics for view split and found that all heuristics failed with insufficient labeled data. The necessary condition of co-training given in [24] suggested that among all potential view splits, the one which enables the most unlabeled instances connect with labeled examples in the combinative graph is preferred; this was empirically verified in [24] and might give inspiration to develop sound practical view split approaches.

Most previous studies on multi-view learning focused on two views, possibly owing to the fact that less data sets with more than two views are publicly available. With the increasing demand of multimedia data analysis, data with more than two views become more accessible. Extending two-view approaches to more views, however, is not trivial. This is because helpful information are concealed in the relations between the views, while the relations become more complicated with more views. A simple "view-invariant" approach is to train one learner from each view, and then let the learners exploit unlabeled data through the strategy of *majority teach minority*. This strategy has been found effective in single-view multi-learner semi-supervised learning approaches tri-training [26] and co-forest [13], and is expected to be helpful on multi-view data. Furthermore, such a semi-supervised learning process would be easy to combine with committee-based active learning approaches.

5 Conclusion

This article briefly reviews some recent advances in exploiting unlabeled data with multiple views. Now it is known that multi-view is not really needed for disagreement-based semi-supervised learning approaches such as co-training; however, given adequate multiple views, amazing performances such as semi-supervised learning with a single labeled example becomes possible. Multi-view also enables exponential improvement of sample complexity for non-realizable active learning with unbounded Tsybakov noise. Overall, multi-view brings great potential of interesting new findings and strong learning approaches for exploiting unlabeled data.

Acknowledgments. This article summarizes the keynote given at the IAPR Workshop on Partially Supervised Learning (PSL), Ulm, Germany, in September 2011. The author was supported by the National Fundamental Research Program (2010CB327903) and the National Science Foundation of China (61073097, 61021062).

References

1. Abney, S.: Bootstrapping. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 360–367 (2002)
2. Balcan, M.-F., Blum, A., Yang, K.: Co-training and expansion: Towards bridging theory and practice. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17, pp. 89–96. MIT Press, Cambridge, MA (2005)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, pp. 92–100 (1998)
4. Castro, R.M., Nowak, R.D.: Minimax bounds for active learning. IEEE Transactions on Information Theory 54(5), 2339–2353 (2008)
5. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
6. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 208–215 (2008)
7. Dasgupta, S., Littman, M., McAllester, D.: PAC generalization bounds for co-training. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems, vol. 14, pp. 375–382. MIT Press, Cambridge, MA (2002)
8. Du, J., Ling, C.X., Zhou, Z.-H.: When does co-training work in real data? IEEE Transactions on Knowledge and Data Engineering 23(5), 788–799 (2010)
9. Goldman, S., Zhou, Y.: Enhancing supervised learning with unlabeled data. In: Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA, pp. 327–334 (2000)
10. Guo, Q., Chen, T., Chen, Y., Zhou, Z.-H., Hu, W., Xu, Z.: Effective and efficient microprocessor design space exploration using unlabeled design configurations. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, pp. 1671–1677 (2011)
11. Huang, S.-J., Jin, R., Zhou, Z.-H.: Active learning by querying informative and representative examples. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 892–900. MIT Press, Cambridge, MA (2010)
12. Kääriäinen, M.: Active learning in the non-realizable case. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, pp. 63–77 (2006)
13. Li, M., Zhou, Z.-H.: Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans 37(6), 1088–1098 (2007)
14. Muslea, I., Minton, S., Knoblock, C.A.: Selective sampling with redundant views. In: Proceedings of the 17th National Conference on Artificial Intelligence, Austin, TX, pp. 621–626 (2000)

15. Muslea, I., Minton, S., Knoblock, C.A.: Active + semi-supervised learning = robust multi-view learning. In: Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia, pp. 435–442 (2002)
16. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: Proceedings of the 9th ACM International Conference on Information and Knowledge Management, Washington, DC, pp. 86–93 (2000)
17. Settles, B.: Active learning literature survey. Technical Report 1648, Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI (2009), http://pages.cs.wisc.edu/~bsettles/pub/settles.activelearning.pdf
18. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proceedings of the 9th ACM International Conference on Multimedia, Ottawa, Canada, pp. 107–118 (2001)
19. Tsybakov, A.: Optimal aggregation of classifiers in statistical learning. Annals of Statistics 32(1), 135–166 (2004)
20. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
21. Wang, W., Zhou, Z.-H.: Analyzing Co-training Style Algorithms. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 454–465. Springer, Heidelberg (2007)
22. Wang, W., Zhou, Z.-H.: On multi-view active learning and the combination with semi-supervised learning. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 1152–1159 (2008)
23. Wang, W., Zhou, Z.-H.: Multi-view active learning in the non-realizable case. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 2388–2396. MIT Press, Cambridge, MA (2010)
24. Wang, W., Zhou, Z.-H.: A new analysis of co-training. In: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, pp. 1135–1142 (2010)
25. Zhou, Z.-H., Chen, K.-J., Dai, H.-B.: Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems 24(2), 219–244 (2006)
26. Zhou, Z.-H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering 17(11), 1529–1541 (2005)
27. Zhou, Z.-H., Li, M.: Semi-supervised regression with co-training style algorithms. IEEE Transactions on Knowledge and Data Engineering 19(11), 1479–1493 (2007)
28. Zhou, Z.-H., Li, M.: Semi-supervised learning by disagreement. Knowledge and Information Systems 24(3), 415–439 (2010)
29. Zhou, Z.-H., Zhan, D.-C., Yang, Q.: Semi-supervised learning with very few labeled training examples. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence, Vancouver, Canada, pp. 675–680 (2007)
30. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI (2006), http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf