Learning with Unlabeled Data and Its Application to Image Retrieval

Zhi-Hua Zhou

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

Abstract. In many practical machine learning or data mining applications, unlabeled training examples are readily available but labeled ones are fairly expensive to obtain because labeling the examples require human effort. So, learning with unlabeled data has attracted much attention during the past few years. This paper shows that how such techniques can be helpful in a difficult task, content-based image retrieval, for improving the retrieval performance by exploiting images existing in the database.

1 Learning with Unlabeled Data

In the traditional setting of supervised learning, a large amount of training examples should be available for building a model with good generalization ability. It is noteworthy that these training examples should be *labeled*, that is, the ground-truth labels of them are known to the learner. Unfortunately, in many practical machine learning or data mining applications such as web page classification, although a large number of unlabeled training examples can be easily collected, only a limited number of labeled training examples are available since obtaining the labels require human effort. So, exploiting unlabeled data to help supervised learning has become a hot topic during the past few years.

Currently there are three main paradigms for learning with unlabeled data, i.e., *semi-supervised learning*, *transductive learning* and *active learning*.

Semi-supervised learning deals with methods for automatically exploiting unlabeled data in addition to labeled data to improve learning performance. That is, the exploitation of unlabeled data does not need human intervene. Here the key is to use the unlabeled data to help estimate the data distribution. For example, a lot of approaches consider the contribution of the unlabeled examples by using a generative model for the classifier and employing EM to model the label estimation or parameter estimation process [7, 9, 12]. Note that previous research on semi-supervised learning mainly focus on classification, while semi-supervised regression only has been studied recently [18]. A recent comprehensive review on semi-supervised learning can be found in [21].

Transductive learning is a cousin of semi-supervised learning, which also tries to exploit unlabeled data automatically. The main difference lies in the different

assumptions they hold: transductive learning assumes that the goal is to optimize the generalization ability on only a given test data set, and the unlabeled examples are exactly the test examples [5, 14]; semi-supervised learning does not assume a known test set, and the unlabeled examples are not needed to be test examples.

Active learning deals with methods that assume the learner has some control over the input space. In exploiting unlabeled data, it requires that there is an *oracle*, such as a human expert, can be queried for labels of specific instances, with the goal of minimizing the number of queries needed. Here the key is to select the unlabeled example on which the labeling will convey the most helpful information for the learner. There are two major schemes, i.e. *uncertainty sampling* and *committee-based sampling*. Approaches of the former train a single learner and then query the unlabeled examples on which the learner is least confident [6]; while approaches of the latter generate a committee of multiple learners and select the unlabeled examples on which the committee members disagree the most [1, 11].

2 A Machine Learning View of CBIR

With the rapid increase in the volume of digital image collections, content-based image retrieval (CBIR) has attracted a lot of research interests [13]. The user can pose an example image, i.e. user query, and ask the system to bring out relevant images from the database. A main difficulty here is the gap between high-level semantics and low-level image features, due to the rich content but subjective semantics of an image. *Relevance feedback* has been shown as a powerful tool for bridging this gap [10, 15]. In relevance feedback, the user has the option of labeling a few images according to whether they are relevant to the target or not. The labeled images are then given to the CBIR system as complementary queries so that more images relevant to the user query can be retrieved from the database.

In fact, the retrieval engine of a CBIR system can be regarded as a machine learning process, which attempts to train a learner to classify the images in the database as two classes, i.e. positive (relevant) or negative (irrelevant). However, this learning task has something different from traditional supervised learning tasks, which makes it interesting and challenging.

First, few users will be so patient to provide a lot of example images in the retrieval process. Therefore, even with relevance feedback, the number of labeled training examples are still very small. Second, few users will be so patient to take part in a time-consuming interaction process. Therefore, the learning process should meet the real-time requirement. Third, instead of returning a crisp binary classification, the learner is expected to produce a rank of the images. The higher the rank, the more relevant the image. Fourth, in typical supervised learning the concept classes are known in advance, but in CBIR, since an image can be relevant to one query but irrelevant to another, the concept classes are dynamic, cannot be given a priori. The last but not least important, typical

machine learning algorithms regard the positive and negative examples interchangeably and assume that both sets are distributed approximately equally. In CBIR although it is reasonable to assume that all the positive examples belong to the same target class, it is usually not valid to make the same assumption for the negative ones because different negative examples may belong to different irrelevant classes and the small number of negative examples can hardly be representative for all the irrelevant classes.

3 Why Exploiting Images in Database?

Section 2 mentioned that in CBIR, even with relevance feedback, the number of example images provided by the user is still very limited. However, there are abundant images existing in the database. Can those images be helpful? Of course.

It is well-known that a main difficulty of CBIR is the gap between high-level semantics and low-level image features, due to the rich content but subjective semantics of an image. This problem can hardly be solved by simply using stronger visual features, but can be released to some degree by using more example images. Usually, the target concept being queried by the user becomes more clear when the user gives more example images. In fact, the relevance feedback mechanism works simply because more example images are given by the user during the feedback process.

Thus, considering the example images as labeled training examples and the images in the database as unlabeled training examples, the CBIR problem resembles what has motivated the research on learning with unlabeled examples. That is, there are a limited number of labeled training examples which are not sufficient for training a strong learner, but there are abundant unlabeled training examples which can be exploited. So, it is evident that techniques of learning with unlabeled data can be used to help improve the retrieval performance.

Note that when the CBIR process is executed on a given database, the task can be mapped to a transductive learning problem since the generalization ability on the given database is concerned; when the CBIR process is executed on an open image source, such as the web, the task can be mapped to a semi-supervised learning problem. On the other hand, since relevance feedback involves human interaction, active learning can be helpful. Thus, CBIR provides a good arena for techniques of learning with unlabeled data.

4 Some Results

We have designed some *co-training* style techniques for exploiting unlabeled data in CBIR [16, 17].

Co-training was proposed by Blum and Mitchell [2], which has then been studied and extended by many researchers and thus become a popular scheme in learning with unlabeled data. In its original version, co-training trains two classifiers separately on two *sufficient and redundant views*, i.e. two attribute

sets each of which is sufficient for learning and conditionally independent of the other given the class label, and uses the predictions of each classifier on unlabeled examples to augment the training set of the other. Later, variants which do not require sufficient and redundant views have been presented [3, 19], and so does an active learning variant [8].

In order to avoid a complicated learning process such that the real-time requirement of CBIR can be met, we [16, 17] employ a very simple model to realize two learners which use Minkowski distances with different orders to measure the image similarities. Each learner will give every unlabeled image a rank which is a value between -1 and $+1$, where positive/negative means the learner judges the concerned image to be relevant/irrelevant, and the bigger the absolute value of the rank, the stronger the confidence of the learner on its judgement. Then, each learner will choose some unlabeled images to label for the other learner according to the rank information. After that, both the learners are re-trained with the enlarged labeled training sets and each of them will produce a new rank for the unlabeled images. The new ranks generated by the learners can be easily combined, which results in the final rank for every unlabeled image. Then, unlabeled images with top ranks are returned as the retrieval results which are displayed according to descending order of the real value of their ranks. Besides, unlabeled images with bottom absolute ranks (i.e. near 0) are put into a pool, which is then used for the user to give feedback. By using such an active learning scheme, the images labeled by the user in the relevance feedback process can have bigger chance to be the ones that are most helpful in improving the retrieval performance. It has been shown that introducing both semi-supervised learning and active learning into CBIR are beneficial [16, 17].

The above approach works with the relevance feedback process, where there are several labeled training examples that can be used. As for the initial retrieval, since there is only one labeled training example, i.e. the user query, exploiting unlabeled examples is more difficult. Such an extreme setting has not been studied before in the area of learning with unlabeled data. In a recent work [20] we have shown that when the images are with textual annotations, even in the initial retrieval, exploiting unlabeled images to improve the retrieval performance is still feasible. The key is to induce some additional labeled training examples by using Kernel Canonical Component Analysis [4] to exploit the correlations between the visual features and textual annotations. Such an approach can be easily generalized to other cases where there is only one labeled training example but the data have two views.

5 Conclusion

Techniques of learning with unlabeled data are helpful in diverse machine learning or data mining applications. This paper shows that how they can be helpful in enhancing the performance of CBIR, which exhibits an encouraging new direction for image retrieval research. Note that here we do not claim that the retrieval performance can be boosted to a level which can 'make the user sat-

isfied', which is still a long way to go. In fact, what we have claimed is that the retrieval performance can be enhanced by exploiting the unlabeled images. Moreover, we believe that CBIR can raise many interesting machine learning research topics, the outputs of which will be not only beneficial to CBIR but also be able to generalize to other learning tasks.

Acknowledgment

This research was partially supported by FANEDD (200343) and NSFC (60325207).

References

1. N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the 15th International Conference on Machine Learning*, pages 1–9, Madison, WI, 1998.
2. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.
3. S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proceedings of the 17th International Conference on Machine Learning*, pages 327–334, San Francisco, CA, 2000.
4. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
5. T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, Bled, Slovenia, 1999.
6. D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, Ireland, 1994.
7. D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In M. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 571–577. MIT Press, Cambridge, MA, 1997.
8. I. Muslea, S. Minton, and C. A. Knoblock. Selective sampling with redundant views. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 621–626, Austin, TX, 2000.
9. K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.
10. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.
11. H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the 5th ACM Workshop on Computational Learning Theory*, pages 287–294, Pittsburgh, PA, 1992.
12. B. Shahshahani and D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994.

13. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
14. V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
15. X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: a comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.
16. Z.-H. Zhou, K.-J. Chen, and H.-B. Dai. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 24(2), 2006.
17. Z.-H. Zhou, K.-J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. In *Proceedings of the 15th European Conference on Machine Learning*, pages 525–536, Pisa, Italy, 2004.
18. Z.-H. Zhou and M. Li. Semi-supervised learning with co-training. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 908–913, Edinburgh, Scotland, 2005.
19. Z.-H. Zhou and M. Li. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005.
20. Z.-H. Zhou, D.-C. Zhan, and Q. Yang. Semi-supervised learning with a single labeled example. *Unpublished manuscript*.
21. X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2005. http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf.