A Survey of Semi-supervised Learning Research

Yapei Zhao

Abstract— **With the rapid growth of modern data sets and increasingly passive data collection, the cost of tagged data is becoming higher and higher, and unlabeled data is not only cheap but sufficient in many applications. The use of unlabeled data to improve the prediction of machine learning systems is a semi-supervised learning problem (SSL). Semi-supervised learning has high practical application value and can improve the learning performance of the model. This article will describe the semi-supervised learning method in three parts. Firstly, it describes the definition and development process of semi-supervised learning, then describes the commonly used algorithms and practical applications of semi-supervised learning, and finally discusses the future research direction of semi-supervised learning.**

***Index Terms*—semi-supervised learning, tagged data; unmarked data; algorithms.**

I. INTRODUCTION

Supervised learning is to use the existing labeled training set for training and learning. Unsupervised learning is to let the computer learn by itself using unlabeled sample sets, such as clustering algorithms. However, with the explosive growth of data sets, it is becoming increasingly difficult to collect large amounts of labeled data. For supervised learning, if the data set of labeled data is too small, the training model effect will not be very good, and the test results are even more horrible. For unsupervised learning, if only a large number of unlabeled data sets are used, the value of labeled sample sets may be lost. Therefore, Merz et al. proposed the concept of semi-supervised learning in 1992. Semi-supervised learning (SSL) refers to learning from both labeled and unlabeled data. Unlike supervised learning that only uses labeled data for training, it does not require a large amount of labeled data, reduces labor costs, and is unlabeled The data is easier to obtain, and the data set is also very sufficient, which makes the training model have better generalization ability and obtain higher accuracy. Because of the above advantages, people are paying more and more attention to semi-supervised learning, and apply semi-supervised learning to various fields, such as semi-supervised learning classification based on SAR images, semi-supervised deep learning brain tissue segmentation, etc.

Since Merz et al. proposed the concept of semi-supervised learning in 1992, more and more people have realized the value of unlabeled sample sets. They have begun to use semi-supervised learning to conduct experiments, and have achieved unexpected results. Semi-supervised learning has been applied to various fields. Here are a few examples to briefly illustrate the development of semi-supervised learning.

In 2013, "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks"[1] This paper proposes a simple and effective semi-supervised learning method for deep neural networks. Basically, the proposed network is trained in a supervised manner, with both labeled and unlabeled data. For unlabeled data, use Pseudo-Labels as ground truth to train the next network, that is, use Unlabeled data to go through the network first to get the pesudo label, and then use these pesudo labels as GT for supervised learning, and the loss function The form is also the addition of each part with a balance coefficient. Just pick the classes with the largest predicted probability as if they were real labels.

In 2015, the paper "Semi-supervised Learning with Ladder Network" [2] proposed a ladder network. In a deep neural network, the ladder network combines supervised learning and unsupervised learning, and adds a classifier at the highest level. , The ladder network becomes a semi-supervised model. Generally, unsupervised learning is only used for network pre-training, followed by normal supervised learning. In the ladder network, the sum of the supervised and unsupervised cost functions is minimized by backpropagation, avoiding the need for hierarchical pre-training. The ladder network can achieve the most advanced performance in semi-supervised MNIST and CIFAR-10 classification, and the arrangement of all labels is unchanged in MNIST classification.

The hottest breakthrough in semi-supervised learning can be said to be the combination with GAN. In 2016, "Improved Techniques for Training GANs" published on nips [3], like many deep generative models, apply GAN to semi-supervised learning, but they To continue and improve this work, improve the effectiveness of network generation confrontation for semi-supervised learning, and achieve the most advanced results in semi-supervised classification.

In short, the progress of semi-supervised learning is quite large. OpenAI uses GAN and 100 pieces of labeled data to do semi-supervised learning, and the accuracy of handwritten digit recognition is consistent with the accuracy obtained by training with 65,000 pieces of labeled data. This technology has been used by Stanford University to detect the distribution of famine in Africa. Only 5% of the satellite maps used for training data are labeled supervised learning data, and the remaining 95% of satellite map data are unlabeled.

II. COMMON ALGORITHMS AND APPLICATION SCENARIOS

Semi-supervised learning according to different classification standards, the classification of the algorithm is also different.

Manuscript received October 06, 2020.
Yapei Zhao, School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin 300387, China

Here, semi-supervised learning is divided into 4 categories according to learning purposes: semi-supervised clustering, semi-supervised classification, semi-supervised regression, and semi-supervised dimensionality reduction.

A. *Semi-supervised Clustering*

Semi-supervised clustering algorithm is based on unsupervised clustering, and studies how to use a small amount of supervised information to improve clustering performance. A small amount of supervision information can be labeled data or constraint relationships between data. Semi-supervised clustering algorithms mainly fall into the following two categories: distance-based methods and large interval methods. For example, the common K-means algorithm is a typical distance-based clustering algorithm. kmeans usually uses Euclidean distance as a metric. But sometimes only the unsupervised K-means algorithm is used. Because of the limitations of the K-means algorithm, a good generalization model cannot be obtained. Therefore, people try to use semi-supervised clustering algorithms to improve the accuracy of clustering methods.

B. *Semi-supervised Classification*

The main idea of semi-supervised classification is to start from the perspective of supervised learning. When the labeled sample set is insufficient, a large number of unlabeled sample sets are used to assist the training of the classifier. There are many common semi-supervised classification methods, including generative model parameter estimation methods, difference-based methods, graph cutting methods, and collaborative training methods. For example, graph-based semi-supervised classification methods, graph construction is a basic component of graph-based semi-supervised learning, which can reveal the geometric structure of manifolds. However, its data density distribution and label information are easily overlooked, and its scalability is relatively poor. In order to solve this problem, Li et al. [4] proposed to adopt a variant of the diffusion process, called self-enhanced diffusion, which can use label information. As for the data density distribution, an intuitive affinity word is introduced, called self-affinity, which can approximate the density distribution well and can be directly spread on the graph. Extensive experiments on noisy synthetic data and various real-world data have proved the effectiveness of this method.

C. *Semi-supervised Regression*

Regression is a statistical analysis method that studies the relationship between two sets of variables. For example, input a training sample of various data of a person into the model, and produce the result of "inputting a person's data and judging the person's economic ability in the future 20 years from now." The result is continuous, and a regression curve is often obtained. When the input independent variables are different, the output dependent variable is not discretely distributed. The purpose of semi-supervised regression and semi-supervised classification is roughly similar, but the sample set in semi-supervised regression is real-valued output and is very smooth. For semi-supervised regression, the clustering assumption is generally invalid, and the manifold assumption is still valid. Common semi-supervised regression methods include difference-based methods and manifold learning methods. The following will describe the algorithm of semi-supervised regression with specific examples.

In 2005, Zhou proposed a semi-supervised regression method GOREG based on co-training. The algorithm does not require too many views, but uses different parameters of the same learner to set and generate two initial learners. After that, they [ZhouL07] extended COREG to use different distance measures, different numbers of neighbors, and other regression models. CMU researcher Yichong Xu et al. proposed a semi-supervised ranking regression (Ranking-Regression), which combines a small number of labeled samples and a large number of unlabeled samples to escape the curse of dimensionality. Semi-supervised regression is not only very useful but also very interesting. Dai Liqing of Nanjing University of Science and Technology and others proposed a geometrically beautiful score based on semi-supervised regression learning, which combines manifold learning with semi-supervised learning and analyzes the face through semi-supervised regression. The geometric beauty score is effective and practical.

D. *Semi-supervised Dimensionality Reduction*

In machine learning and other fields, processing of high-dimensional images and videos is not only difficult and time-consuming, but it is also easier to form a dimensional disaster. This requires dimensionality reduction to overcome these problems. Using the idea of semi-supervised learning to reduce dimensionality results in a new branch called semi-supervised dimensionality reduction. Commonly used algorithms for semi-supervised dimensionality reduction include class label-based methods, pair-wise constraint-based methods, and other popular embedding-based methods and sample correlation-based methods. Semi-supervised dimensionality reduction has been applied to various aspects. For example, in the paper "Face Recognition Based on Semi-supervised Dimensionality Reduction" [5], the semi-supervised dimensionality reduction method based on paired constraints is used for face recognition. Experiments have proved that the face recognition rate is higher than the ordinary principal component analysis dimensionality reduction method.

III. FUTURE RESEARCH DIRECTIONS

With the efforts of many scientific researchers, the algorithms of semi-supervised learning have become more and more abundant, and the research of semi-supervised learning has also made certain achievements, which are applied to different areas of life. But this is only preliminary research, and there are still some practical problems to be explored.

The anti-interference performance of semi-supervised learning is relatively poor. For example, I tried to use semi-supervised learning to classify SAR images, but because most SAR images are noisy, the final classification effect is not very satisfactory. The results are shown in Table 1. Show. Later, after consulting the information, it is known that most semi-supervised learning is based on the premise that the data used is noise-free. However, in practical applications, it is difficult to obtain a data set without noise interference.

Therefore, choosing a suitable SSL method

Table 1 SAR image classification based on semi-supervised learning

	Number of images in each category	The number of correctly classified images of each type	Accuracy
Airplane	171	120	
Baseball field	180	104	
Beach	179	89	
Farmland	178	122	
Forest	180	102	
Residential	179	120	
Parking lot	180	128	
Port	180	104	
Total	1427	889	62.3%

based on actual problems, and making better use of examples of classless labels to help improve the accuracy and speed of learning is a future research direction.

In fact, the reason for the unsatisfactory experimental effect is not only noise interference. The selection of training examples and parameters is also a key factor. In most semi-supervised learning experiments, a semi-supervised learning method is manually selected and learning parameters are set. At the same time, the performance of semi-supervised learning is better than that of supervised learning and unsupervised learning, but when the selected method does not match the task Or when the parameter settings are not appropriate, the performance of semi-supervised learning will be worse than that of supervised learning or unsupervised learning. Therefore, matching the appropriate semi-supervised learning method according to the learning task and setting the appropriate parameters is the content of the future semi-supervised learning that needs to be studied in depth.

IV. CONCLUSION

At this stage, semi-supervised learning is mostly an extension of supervised learning. If the model is incorrect or the parameters are not properly selected, the performance of semi-supervised learning may be lower than that of the same type of supervised learning. For neural networks, a good initialization can make the results more stable, with fewer iterations. Therefore, how to use untagged data to make the network have a good initialization has become a research point. In general, a single method should not be enough, because not all samples can be classified well. If classification is difficult, how to perform semi-supervised classification. In short, the semi-supervised learning method has achieved certain results, and we believe that it will still be a hot issue for a period of time in the future, with a wide range of application prospects.

REFERENCES

[1] Dong-Hyun Lee.Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks[C].Korea:Nangman Computing, 2013
[2] Antti Rasmus*.Semi-Supervised Learning with Ladder Networks[C].nips,2015
[3] Tim Salimans,Ian Goodfellow.Improved Techniques for Training GANs[C].Machine Learning,2016
[4] Qilin Li∗ , Wanquan Liu, Lin Li.Self-reinforced diffusion for graph-based semi-supervised learning[J] . Machine Learning , 2019, 39: 103-134.
[5] 陈丽霞.基于半监督降维的人脸识别[D].河北 : 河北大学, 2014
[6] ZHU X J. Semi-supervised Learning Literature Survey [R]. Madison : University of Wisconsin, 2008.
[7] 周志华. 半监督学习中的协同训练算法 [M]// 周志华 , 王珏. 机器学习及其应用. 北京 : 清华大学出版社 , 2007: 259-275.
[8] 苏金树 , 张博锋 , 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报 , 2006, 17(9): 1848-1859.
[9] NIGAM K, MCCAILuM A K, THRUN S, MITCHELI T. Text Classification from Labeled and Unlabeled Documents using EM[J] . Machine Learning, 2000, 39: 103-134.
[10] ZHOU Z H, CHEN K J, YUAN J. Exploiting Unlabeled Data in Content-Based Image Retrieval[C]//Proceedings of the 15th European Conference 0n Machine Learning, Pisa, Italy, 2004: 525-536.
[11] ZHOU z H, CHEN K J, DAI H B. Enhancing Relevance Feedback in Image Retrieval using Unlabeled Data[J]. ACM Transactions on Information Systems, 2006, 24(2): 219-244.
[12] TSUDA K, RATSCH G. Image Reconstruction by Linear Programming[J]. IEEE Transactions on Image Processing, 2005, 14(6): 737-744.
[13] SONG Y Q, ZHANG C S, LEE JG, eta1 . Semi-supervised Discriminative Classification with Application toTumorous Tissues Segmentation of MRLBrain Images[J] . Pattern Analysis& Applications, 2009, 12: 99-115.
[14] TANG J , HUA X S , QI G J , et a1 . Structure-Sensitive Manifold Ranking for Video Concept Detection[C]//Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 2007: 852-861.
[15] YANR , NAPHADE M R . Semi-supervised Cross Feature Learning for Semantic Concept Detectionin Videos[c]//Proceedings of the IEEE Computer Society International Conference on Computer Vision and Pattern Recognition , San Diego, USA , 2005, 1: 657663.
[16] HEJR , LIM J , ZHANG H J , eta1 . Manifold-Ranking Based Image Retrieval[C]//Proceedings of the12th Annual ACM International Conferenceon Multimedia, New York, USA, 2004 :9-16.
[17] FENG W , XIE L , ZENG J , LIU Z Q . Audio-visual Human Recognition using Semi—supervised SpectralLearning and Hidden Markov Models[J]. Journal of Visual Languages and Computing, 2009, 20: 188-195.
[18] SHAHSHAHANI B, LANDGREBE D . The Effect of Unlabeled Samplesin Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon [J] . IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 10871095.