Cross Domain Semi-Supervised Learning using Feature Formulation

Xingquan Zhu*

Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, NSW 2007, Australia
xqzhu@it.uts.edu.au

Extended abstract:
Semi-Supervised Learning (SSL), represents a class of machine learning techniques that utilize unlabeled samples to boost the learning on a labeled set (referred to as the "target set" in this paper). Intuitively, as labeling training samples is often subject to a significant amount of human labor or costs, whereas cheap unlabeled samples can be easily collected with trivial efforts, it makes sense to utilize unlabeled samples in the learning process to boost the model performance. Traditionally, SSL problems are solved through an iterative labeling and learning process, where some automated "labeling agents" trained from labeled samples are used to generate class labels for unlabeled samples, with aggregated training set (containing both genuinely and automatically labeled samples) being used to further improve the labeling agents. The iterative labeling and learning process normally repeats a number of times until some stopping criteria are satisfied. In some situations, all unlabeled samples are included into the training set with a pseudo-class label assigned to each unlabeled sample. The final models trained from the aggregated training set are used to predict future samples.

Many methods exist for semi-supervised learning by using mechanisms, such as Expectation Maximization (EM) principles, graph-based label propagation, or orthogonal neighborhood-preserving projection, to determine proper class labels for unlabeled samples. All these methods, in a narrow sense, share a striking similarity in their design: including unlabeled samples into the training set by assigning a class label to each of them. As a result, unlabeled samples can be directly integrated into the training process in the original feature space. In this paper, we call this approach "primitive semi-supervised learning" (pSSL) mainly because unlabeled samples are included into the training set in a primitive instance form (*i.e.* original feature space). Alternatively, if an unlabeled samples is linked to the training process through some transformed feature space, we call such approaches formative semi-supervised learning (fSSL). Under this definition, most existing algorithms, such as Co-Training, common component, ASSEMBLE, SemiBoost, and Transductive SVM all belong to pSSL because they explore the connection between labeled and unlabeled samples in the original feature space.

Traditional pSSL approaches, in practice, suffers from a number of disadvantages including false labeling and incapable of utilizing out-of-domain samples. In this paper, we propose a formative Semi-Supervised Learning (fSSL) framework which explores hidden features between labeled and unlabeled samples to achieve semi-supervised learning. fSSL regards that both labeled and unlabeled samples are generated from some hidden concepts with labeling information partially observable for some samples. The

* corresponding author. Tel: +61-2-9514-1885

key of the fSSL is to recover the hidden concepts, and take them as new features to link labeled and unlabeled samples for semi-supervised learning. Because unlabeled samples are only used to generate new features, but not to be explicitly included in the training set like pSSL does, fSSL overcomes the inherent disadvantages of the traditional pSSL methods, especially for samples not within the same domain as the labeled instances. The inherent advantage of fSSL, in comparison with pSSL, is twofold.

- False Labeling Immunization: fSSL uses unlabeled samples to formulate new features for semi-supervised learning. As a result, there is no risk of including mislabeled samples into the training set. Although new features may also contain erroneous values (or some noninformative values from the learning task perspective), since attribute errors are essentially less harmful than class label errors, fSSL is much less vulnerable to data errors and has a much lower risk of performance loss than pSSL.
- Cross-Domain Learning: fSSL does not require the assignment of a class label for each unlabeled sample (whereas most pSSL methods do). As a result, even if unlabeled samples are from different domains of the labeled samples, fSSL is still able to integrate them into the training process to boost the learning on the target set.

Experimental results and comparisons demonstrate that fSSL significantly outperforms pSSL based methods for both within-domain and cross-domain semi-supervised learning.

Keywords: Machine Learning, Classification, Semi-Supervised Learning, Cross Domain Learning.