Label Propagation (to be reviewed)

A. Tsanakas, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 1.1

Artificial Intelligence & Information Analysis Lab

Label Propagation

- **Introduction**
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

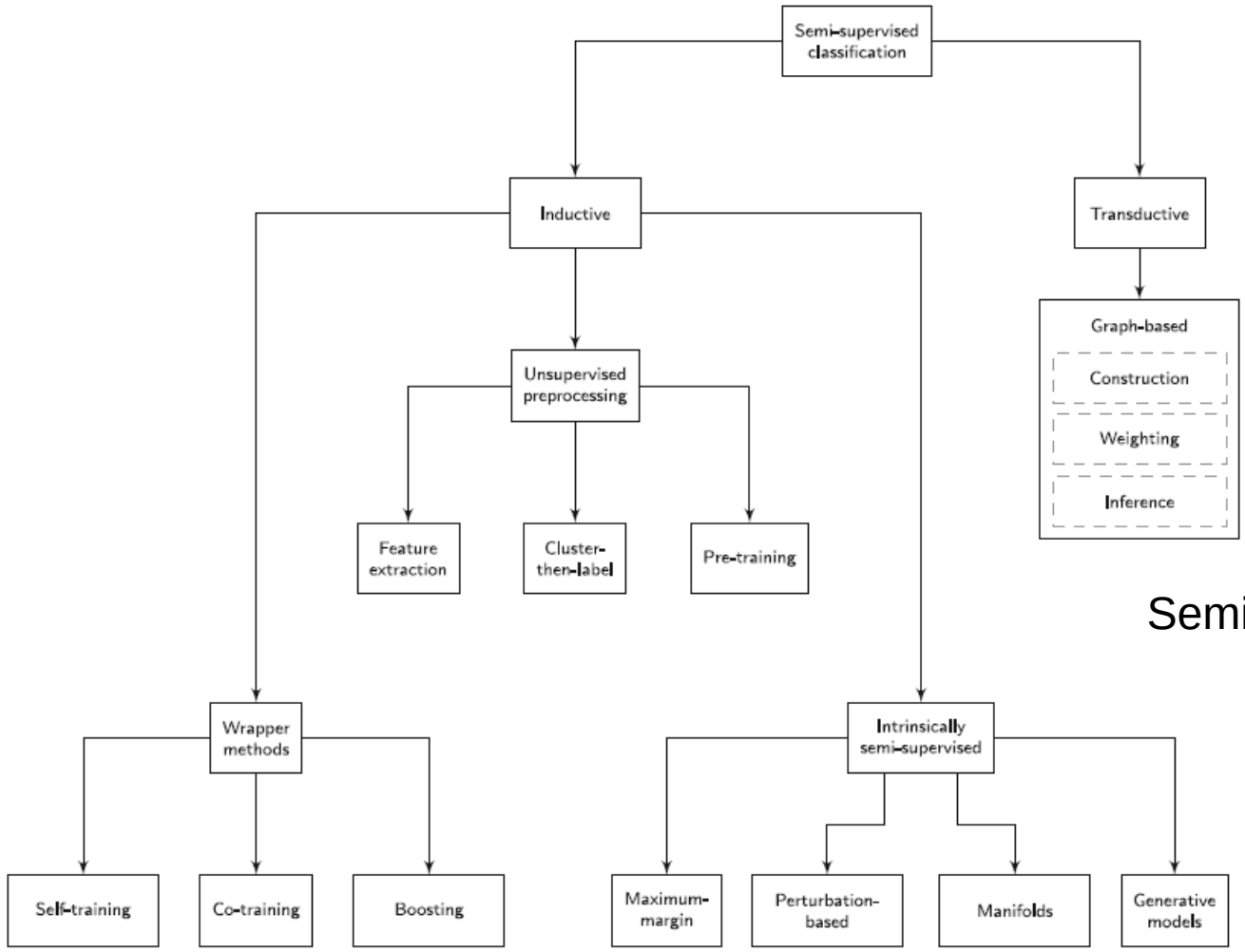
Introduction

- The two most important tasks in machine learning are traditionally supervised and unsupervised learning.
  - ✓ In supervised learning, we have a set of data points consisting of some input x and a label value y for each input, and we try to construct a classifier or regressor that can estimate the output value for previously invisible inputs.
  - ✓ In unsupervised learning, we don't have an output value y and we try to infer some underlying structure from the x inputs improve learning accuracy.

Introduction

- Semi-supervised learning is now one of the major branches of machine learning that seeks to combine two tasks.
  - Semi-supervised learning is a set of learning methods that, at the training stage, combine a small amount of labeled data (y inputs) with a large amount of unlabeled.
  - Using a small amount of labeled data can greatly improve learning accuracy.

Introduction



Semi-supervised learning classification

Van Engelen, Jesper E., and Holger H. Hoos. "A survey on semi-supervised learning." *Machine Learning* 109.2 (2020): 373-440.

Introduction

- Inductive classifiers learn a global representation of the data feature space and can be applied to "unseen" data that were not included in the original dataset.

- On the other hand, transductive semi-supervised classifiers learn a local representation of the data feature space and can be applied in a specific dataset consisting of labeled and unlabeled data.

Introduction

- The most popular transductive learning algorithms are:
  - graph construction approaches
  - label inference methods

- These approaches are also called Label propagation algorithms.

Label Propagation

- Introduction
- **Graph construction approaches**
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Graph Construction Approaches

- In label propagation the similarity graph reflects the similarity between the entities being labeled, with respect to the specific task, and as a result is critical to the performance of the process.

Graph Construction Approaches

- Graph construction methods are divided in 3 categories:

    1. **Adjacency matrix construction**
    2. **Graph weighting**
    3. **Simultaneous graph construction and weighting**

Label Propagation

- Introduction
- Graph construction approaches
  - **Adjacency Matrix Construction**
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Adjacency Matrix Construction

- In these algorithms, the creation of an adjacency table before the construction of the graph plays an important role.
- The elements of the neighborhood matrix indicate the presence of edges between pairs of nodes.
- We present the most important algorithms of this category, which are: ε-neighborhood, k-nearest neighbors and b-matching.

Adjacency Matrix Construction

- **ε-neighbourhood**
- In this approach, x_i and x_j node is connected if the following mathematical property holds:

$$d(x_i, x_j) \leq \varepsilon,$$

where $d(\cdot,\cdot)$ is some distance measure.

- If the scale of patterns varies across the given input data, the method does not give good results, since the ε is fixed.

Adjacency Matrix Construction

- **k-nearest neighbours (k-NN)**
- In k-nearest neighbours (k-NN) graphs, each node is connected to its k nearest neighbors.
- A disadvantage of this approach is that the graphs are asymmetric since certain nodes end up with a higher degree than others.

Adjacency Matrix Construction

- **k-nearest neighbours (k-NN)**
- To solve this problem many works have proposed some additional processing to obtain an undirected graph:

Symmetric k-NN	Constructs an edge if I is in the k- neighborhood of j or vice versa
Mutual k-NN	Constructs an edge if I and j are both in each other's k-neighborhood

Adjacency Matrix Construction

- **b-matching based method**
- A method that ensures that each node has the same number of neighbours, and that the nodes have exactly the requested number of edges.
- They define the distance matrix C as:

$$C_{ij} = \sqrt{\{W_{ii} + W_{ji} - 2W_{ij}\}},$$

these notions are equivalent.

Adjacency Matrix Construction

- **b-matching based method**
- The corresponding optimization problem is formulated as:

$$\min_{A \in \mathbb{B}^{n \times n}} \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} C_{ij}$$

$$Subject\ to \sum_{i=1}^{n} A_{ij} = b\ , i = 1, \dots, n,$$

$$A_{ij} = 0, \qquad i = 1, \dots, n,$$

$$A_{ij} = A_{ji}, \qquad j, i = 1, \dots, n,$$

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - **Graph Weighting**
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Graph Weighting

- The second step of graph construction is to determine the weights for the edges in the graph.

- In many works, proposed for weighting, *the Gaussian edge weighting*:

$$W_{ij} = \exp\left(\frac{\left|x_i - x_j\right|^2}{2\sigma^2}\right)$$

Graph Weighting

- Another algorithm is based on the assumption that the graph must be constructed so that any data point x_j can be accessed as linear combination of its neighbors.
- This algorithm is called *linear neighborhood propagation*(LNP) and the weights are given by the following formula:

$$x_j = \sum_{u_i \in N(u_j)} W_{ji}\, x_i + \varepsilon_j$$

Graph Weighting

- A modification of the symmetric k-nearest neighbour method based on the following rule: two nodes are connected if either of them is in the other's k-neighbourhood, but the weight of the two connections is summed if they are both in each other's neighbourhoods.
- The modified weight matrix W is constructed based n the original weight matrix $\widehat{W}$ as follows:

$$W_{ij} = \begin{cases} \widehat{W_{ij}} + \widehat{W_{ji}} \;\; if \; u_i \in N(u_j) \; and \; u_j \in N(u_i) \\ \widehat{W_{ji}} \;\; if \; u_i \in N(u_j) and \; N \notin N(u_i) \\ \widehat{W_{ij}} \quad otherwise \end{cases}$$

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - **Simultaneous Graph Construction and Weighting**
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Simultaneous Graph Construction and Weighting

- Many works have proposed a different method to infer the weights and the graph structure.

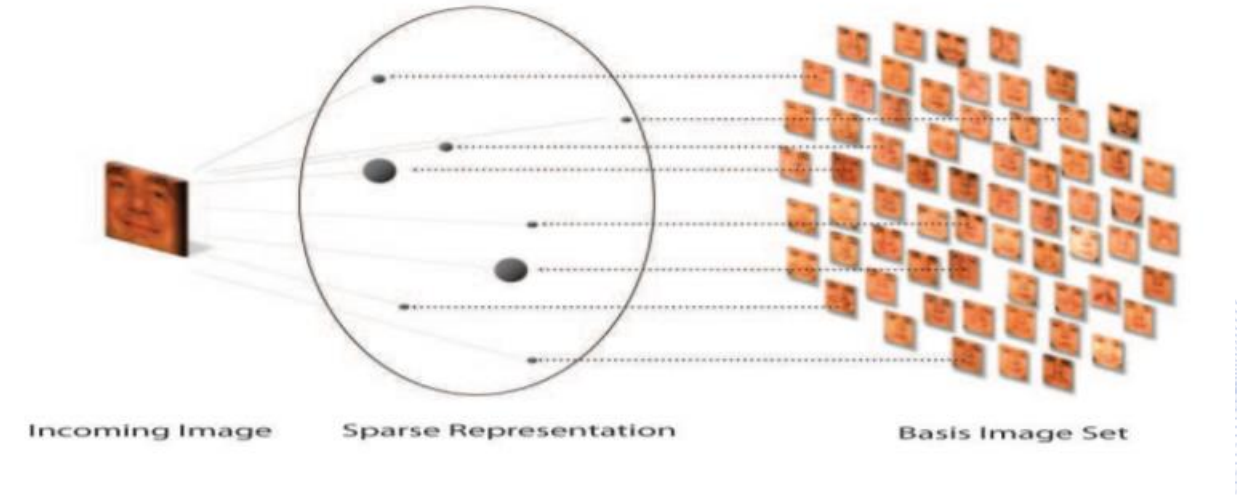
- These methods are based on linearly reconstructing nodes based on all other nodes, in contrast with the LNP algorithm which for each node reconstruct as a linear combination of its neighbours.

Simultaneous Graph Construction and Weighting

- A based on the sparse coding approach formulated for face recognition.
- This method based on the calculation of a coefficient vector $a \in R^n$ for each node which denote the contributions of all other nodes to the re-construction of x_i .
- This reconstruction is then calculated as $x_i = (X')^T a$, where $(X')^T \in R^{nxd}$ denotes the full data matrix, but with a row of zeroes at index i.

Simultaneous Graph Construction and Weighting

- *Sparse representation for the incoming image using basis images*.



Yan, Shuicheng, and Huan Wang. "Semi-supervised learning by sparse rep-resentation." Proceedings of the 2009 SIAM international conference on datamining. Society for Industrial and Applied Mathematics, 2009.

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- **Label Inference Methods**
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Label Inference Methods

- When the similarity graph has been constructed, label propagation is performed on the graph nodes, through a **label inference** method.

- **Label Inference** determines how the labels are spread from the set of labeled nodes to the unlabeled nodes according to the rules that govern label dissemination, and the type and the number of graphs they apply.

Label Inference Methods

- Label inference methods divided in four basic categories:
  - *graph min-cut*
  - Probabilistic label assignments: *Markov random fields*
  - Efficient probabilistic label assignments: *Gaussian random fields*
  - Handling label noise and irregular graphs: *local and global consistency*

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - **Graph Min-cut**
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Graph Min-cut

- The first graph-based semi-supervised classification.
- this approach combines the k-nearest neighbours algorithm and the ε- neighbourhood.
- Specifically, a single *source* node v+ is added and connected within finite weight to the positive data points, and a single *sink* node v−, connected within finite weight to the negative data points.

Graph Min-cut

- The min-cut algorithm based on finding a set of edges with a minimal combined weight that, when removed, result in a graph with no paths from the source node to the sink node.

- All unlabeled nodes that are in the component containing v+ are labelled as positive (+), and all unlabeled nodes that are in the component containing v− are labelled as negative (-), based in the resulting graph from the previous steps.

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - **Markov Random Fields**
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Markov Random Fields

- A basic disadvantage of the min-cut is the lack of an efficient way of estimating classification probabilities.

- It would be important to estimate the probability that an unlabeled data point xi has label c.
- In other words, the probability: $P(y_i = c)$

- The solution to this problem gives an approaching graph-based method from the perspective of *Markov random fields.*

Markov Random Fields

- *Hammersley-Clifford theorem* : A probability distribution $P(X = x)$ (where X and x to denote random variables and their realizations) for random variables $X_1, \ldots, X_n$ corresponds to a Markov random field if a graph G exists such, that the joint probability function $P(X = x)$, can be factorized over the (maximal) cliques of G .

Markov Random Fields

- Set Z is a normalization constant, C_G is the set of cliques in G , ψ_c is an arbitrary function, and x_c contains the realizations of random variables in clique c.
- $P(X = x)$ corresponds to a Markov random field formed by G if:

$$P(X = x) = \frac{1}{Z} \cdot \prod_{c \in C_G} \psi_c(x_c)$$

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - **Gaussian Random Fields**
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Gaussian Random Fields

- There is no closed-form solution for calculating the marginal probabilities in the Markov random field with binary labels described previously.
- However, when the random variables $\hat{Y}$ are relaxed to take real values, a closed-form solution exists.
- It involves fixing the labels of the labelled data points and using quadratic cost for the pairs of predictions $\hat{y}_i, \hat{y}_j \in \mathrm{R}$.
- This results in an objective function identical to that used in the min-cut formulation, except for the relaxation of the predictions to real numbers.

Gaussian Random Fields

- Using real-valued predictions with a quadratic loss function, the exponential form for $P(\hat{Y} = \hat{y})$ is a multivariate Gaussian distribution. Thus, a closed-form solution for the mode of the field, which equals its mean, exists.
- Furthermore, the marginal probability distribution $P(\hat{Y} = \hat{y})$ is Gaussian as well, allowing for computation of the label predictions minimizing the error rate.
- This is why the random field is called a Gaussian random field.

Gaussian Random Fields

- We defined the graph Laplacian as $L = D - W$, where D is the degree matrix (i.e. a diagonal matrix with the vertex degrees on the diagonal)
- The predicted label at each n unlabeled data point is equal to the average of the predictions of its neighbors, i.e

$$\widehat{y_i} = \frac{1}{D_{ii}} \cdot \sum_{u_j \in N(u_i)} W_{ij} \cdot \hat{y}_j \quad for\ i = l+1, \dots, n,$$

where $N(v_i)$ denotes the neighbourhood of node v_i, that is, $N(v_i) = \{v_j : W_{\{i\ j\}} \neq 0\}$.

Gaussian Random Fields

- **Label propagation algorithm**

- It is an iterative algorithm that computes soft label assignments $\widehat{y_i} \in R$ by pushing (propagating) the estimated label at each node to its neighbouring nodes based on the edge weights.

- In other words, the new estimated label at each node is calculated as the weighted sum of the labels of its neighbors. In matrix notation, let A_{ij} denote the transition matrix as follows:

$$A_{ij} = \frac{W_{ij}}{\sum_{u_k \in N(u_i)} W_{ik}}$$

Gaussian Random Fields

- **Label propagation algorithm**

- The label propagation algorithm then consists of two steps, which are repeated until the label assignment $\hat{y}$ converges. Starting with an initial label assignment $\hat{y}$,which is random for the unlabeled data points and equal to the true labels for the labelled data points:

  - Propagate labels from each node to the neighboring nodes: $\hat{y} = A^T \hat{y}$
  - Reset the predictions of the labelled data points to the corresponding true labels

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - **Local and Global Consistency**
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Local and Global Consistency

- The Gaussian random fields method has two disadvantages: the true labels are clamped to the labeled data points and that means that it does not handle label noise well and also, in irregular graphs, the influence of nodes with a high degree is large at many times.

- To solve that issue proposed an approach closely related to the Gaussian random fields method, which is called the *local and global consistency(LGC) method*.

Local and Global Consistency

- The algorithm:

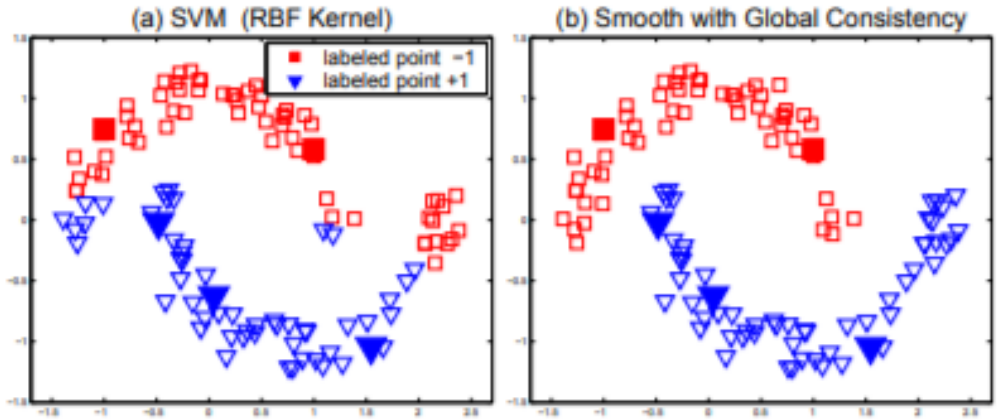
1. Form the affinity matrix W defined by $W_{ij} = \exp\left(\frac{|x_i - x_j|^2}{2\sigma^2}\right)$ and $W_{ij} = 0$.
2. Construct the matrix $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ in which D is a diagonal matrix with its (i, i) −element equal to the sum of the i-th row of W.
3. Iterate $F(t+1) = \alpha S F(t) + (1-\alpha)Y$ until convergence, where is a parameter in $(0, 1)$.
4. Let F^* denote the limit of the sequence $F(t)$ which is equal to: $F^* = \lim_{x \to \infty} F = (I - \alpha S)^{-1} Y$.

5. Label each point xi as a label:

$$y_i = argmax_{j \le c} F_{ij}^*$$

Local and Global Consistency

- Smooth classification results given by supervised classifiers with the global consistency: (a) the classification result given by the SVM with a RBF kernel; (b) smooth the result of the SVM using the consistency method



Zhou, Dengyong, et al. "Learning with local and global consistency." Ad-vances in neural information processing systems 16 (2003): 321-328.

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - **Label Propagation on Data with Multiple Representations**
  - Label Propagation on Hypergraphs
- Label Propagation for Deep Learning

Label Propagation on Data with Multiple Representations

- In real world scenarios, multimedia data is represented in multiple feature spaces.

- In such cases, a separate similarity graph can be constructed for each of these representations.

Label Propagation on Data with Multiple Representations

- The information from the multiple data representations can be fused in two ways. Fusion can take place during graph construction (early fusion) or fusion can be performed at the decision level (late fusion). Late fusion is also referred to as multi-modal fusion or multi-modality learning.

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - **Label Propagation on Hypergraphs**
- Label Propagation for Deep Learning

Label Propagation on Hypergraphs

- Most applications usually exhibit complex non-pairwise relationships, which involve an arbitrary number of data samples such as in the case of multiple labels per sample.

- In order to efficiently represent complex relationships between data samples, hypergraphs are adopted.

Label Propagation on Hypergraphs

- Such methods utilize hypergraph clustering, hypergraph spectral learning, hypergaph normalization, hypergraph Laplacian Regularization for semi-supervised label propagation, random walk interpretation of hypergraph Laplacian Regularization is also presented, and the extension of normalized and ratio cut to hypergraphs.

Label Propagation

- Introduction
- Graph construction approaches
  - Adjacency Matrix Construction
  - Graph Weighting
  - Simultaneous Graph Construction and Weighting
- Label Inference Methods
  - Graph Min-cut
  - Markov Random Fields
  - Gaussian Random Fields
  - Local and Global Consistency
  - Label Propagation on Data with Multiple Representations
  - Label Propagation on Hypergraphs
- **Label Propagation for Deep Learning**

Label Propagation for Deep Learning

- Deep neural networks are used in many vision problems on the computer.

- But obtaining large amounts of annotated training data annotated by humans for every single task is often impossible.

- Iscen, Ahmet, and et al. suggested a novel algorithm that combines transductive learning with modern deep learning.

Label Propagation for Deep Learning

- The Iterative algorithm (Algorithm 2) is defined as follows:
- First, they construct a sparse affinity matrix $A \in R^{n \times n}$ with elements:

$$A_{ij} \doteqdot \begin{cases} [v_i^T v_j]_+^{\gamma}, & if\ i \neq j \bigwedge_i v \in NN_k(v_j) \\ 0, & otherwise \end{cases}$$

where NN_k denotes the set of k nearest neighbors in X, and γ is a parameter following recent work on manifold-based search. Also, let $W = A + A^T$, which is indeed a symmetric nonnegative adjacency matrix with zero diagonal.

Label Propagation for Deep Learning

- Then, calculate the $Z := (I - W)^{-1}Y$and infer the pseudo-labels the class prediction for an unlabeled example xi with the follow way:

$$y_i = argmax_{j \leq c} F_{ij}^*$$

Label Propagation for Deep Learning

- pseudo-labels from matrix Z have two disadvantages: we define pseudo-labels on all unlabeled nodes while we do not have the same certainty for each of them.
- Second, pseudo-labels may not be balanced over classes, which will impede learning. To solve this issue we use entropy, as a measure of uncertainty, to assign weight ω_i to example x_i, defined by:

$$\omega_i = 1 - \frac{H(\widehat{z_{ij}})}{\log(c)}$$

Label Propagation for Deep Learning

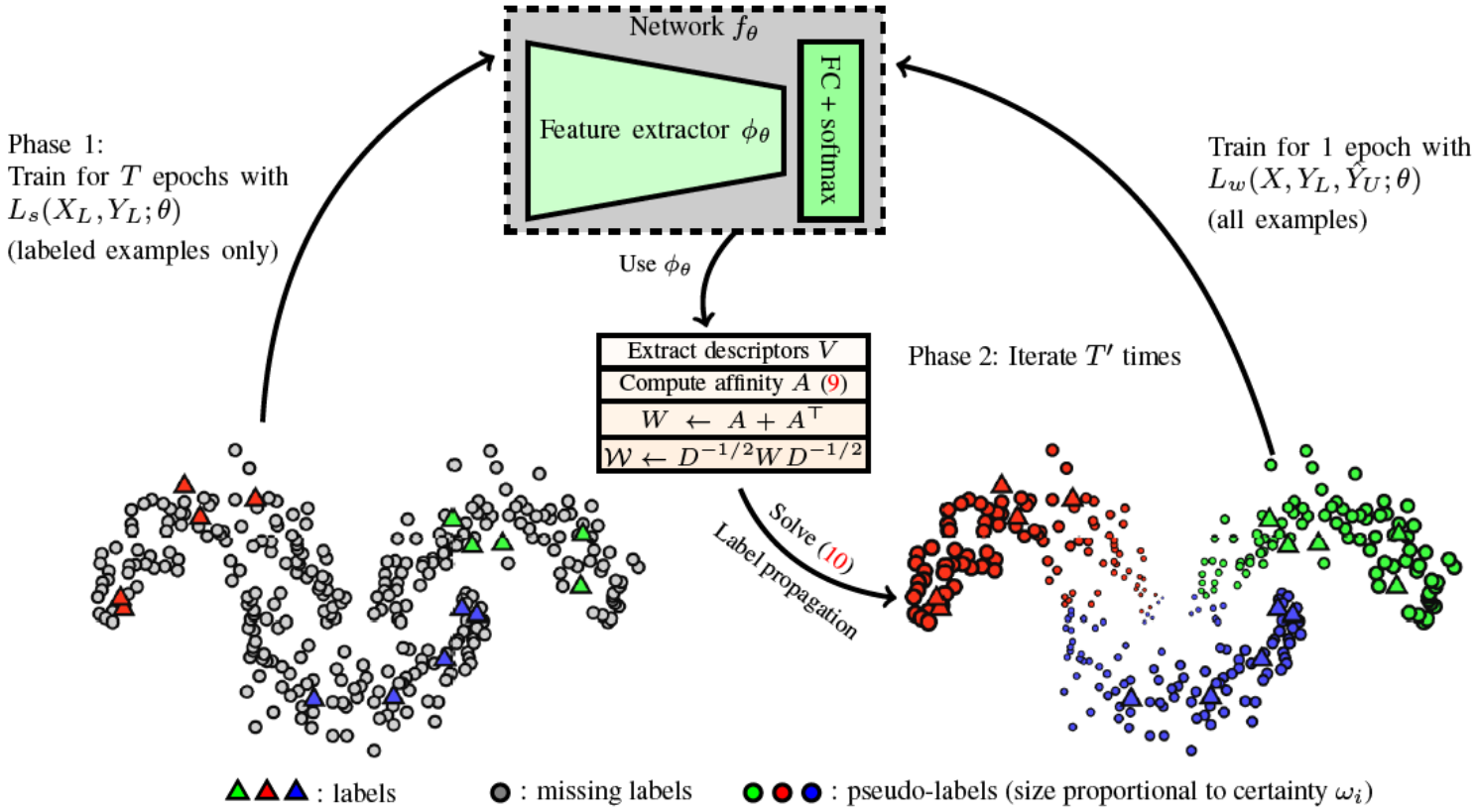
- To deal with the latter issue of class imbalance, we assign weight ζ_j to class j that is inversely proportional to class population, defined as:

$$\mathrm{Z}_j \coloneqq \left(|L_j| + |U_j|\right)^{-1},$$

where L_j(resp.U_j) supervised loss of the examples labeled (resp. pseudo-labeled) as class j.

- Training is performed on all data, using certainty-based weights.

Label Propagation for Deep Learning



Iscen, Ahmet, et al. "Label propagation for deep semi-supervised learning."Proceedings of the IEEE conference on computer vision and pattern recognition.2019.

Bibliography

[PIT2016]I. Pitas, Graph analysis in social media, editor, CRC Press, 2016.

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.

[PIT2017] I. Pitas, “Digital video processing and analysis ” , China Machine Press, 2017 (in Chinese).

[PIT2013] I. Pitas, “Digital Video and Television ” , Createspace/Amazon, 2013.

[NIK2000] N. Nikolaidis and I. Pitas, 3D Image Processing Algorithms, J. Wiley, 2000.

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Bibliography

[ZD03] Zhou, Dengyong, et al. "Learning with local and global consistency." Ad-vances in neural information processing systems 16 (2003): 321-328.[IA19] Iscen, Ahmet, et al. "Label propagation for deep semi-supervised learning."Proceedings of the IEEE conference on computer vision and pattern recognition.2019.

[BAC01] Blum, Avrim, and Shuchi Chawla. "Learning from labeled and unlabeleddata using graph mincuts." (2001).

[VJH20] Van Engelen, Jesper E., and Holger H. Hoos. "A survey on semi-supervisedlearning." Machine Learning 109.2 (2020): 373-440.

Bibliography

ZXG03] Zhu, Xiaojin, Zoubin Ghahramani, and John D. Lafferty. "Semi-supervisedlearning using gaussian fields and harmonic functions." Proceedings of the 20thInternational conference on Machine learning (ICML-03). 2003.[BAS01] Blum, Avrim, and Shuchi Chawla. "Learning from labeled and unlabeleddata using graph mincuts." (2001).

[SARG13] de Sousa, Celso André R., Solange O. Rezende, and Gustavo EAPABatista. "Influence of graph construction on semi-supervised learning." JointEuropean Conference on Machine Learning and Knowledge Discovery inDatabases. Springer, Berlin, Heidelberg, 2013.

[JWC09] Jebara, Tony, Jun Wang, and Shih-Fu Chang. "Graph construction and b-matching for semi-supervised learning." Proceedings of the 26th annual interna-tional conference on machine learning. 2009.

Bibliography

[SCSG13] de Sousa, Celso André R., Solange O. Rezende, and Gustavo EAPA Batista."Influence of graph construction on semi-supervised learning." Joint EuropeanConference on Machine Learning and Knowledge Discovery in Databases.Springer, Berlin, Heidelberg, 2013.

[WJC08] Wang, Jun, Tony Jebara, and Shih-Fu Chang. "Graph transduction via alternating minimization." Proceedings of the 25th international conference onMachine learning. 2008.

[LC09] Liu, Wei, and Shih-Fu Chang. "Robust multi-class transductive learning withgraphs." 2009 IEEE Conference on Computer Vision and Pattern Recognition.IEEE, 2009.

Bibliography

[YW09] Yan, Shuicheng, and Huan Wang. "Semi-supervised learning by sparse rep-resentation." Proceedings of the 2009 SIAM international conference on datamining. Society for Industrial and Applied Mathematics, 2009.9

[ZXZ03] Zhu, Xiaojin, Zoubin Ghahramani, and John D. Lafferty. "Semi-supervisedlearning using gaussian fields and harmonic functions." Proceedings of the 20thInternational conference on Machine learning (ICML-03). 2003.

[ZG02] Zhu, X., Ghahramani, Z. (2002a). Learning from labeled and unlabeled datawith label propagation. Technical Report. CMU-CALD-02-107, Carnegie Mel-lon University.

Q & A

Thank you very much for your attention!

Contact: Prof. I. Pitas
pitas@csd.auth.gr