

Graph-based semi-supervised learning - uzupełnienie

Jerzy Stefanowski

Instytut Informatyki PP

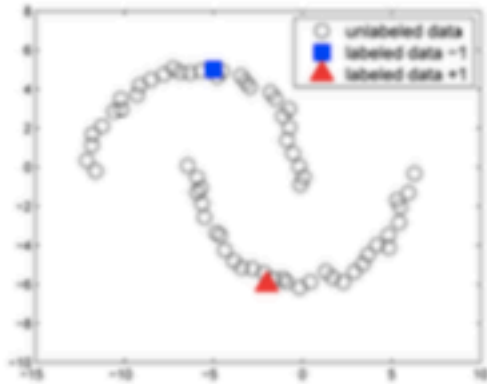
Poznań kwiecień 2020

Ack.

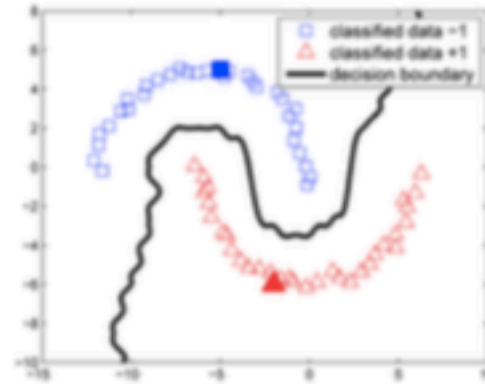
Inspiracje do sajdów tego wykładu

- Zhu J.: Semi-supervised tutorial – slajdy dostępne online
- M.Balcan oraz A.Blum – wykłady o co-training

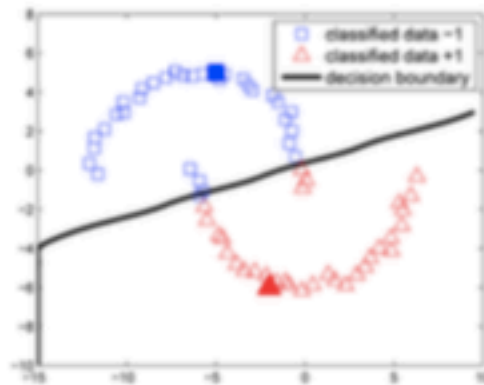
Motywacja podejść grafowych



(a)



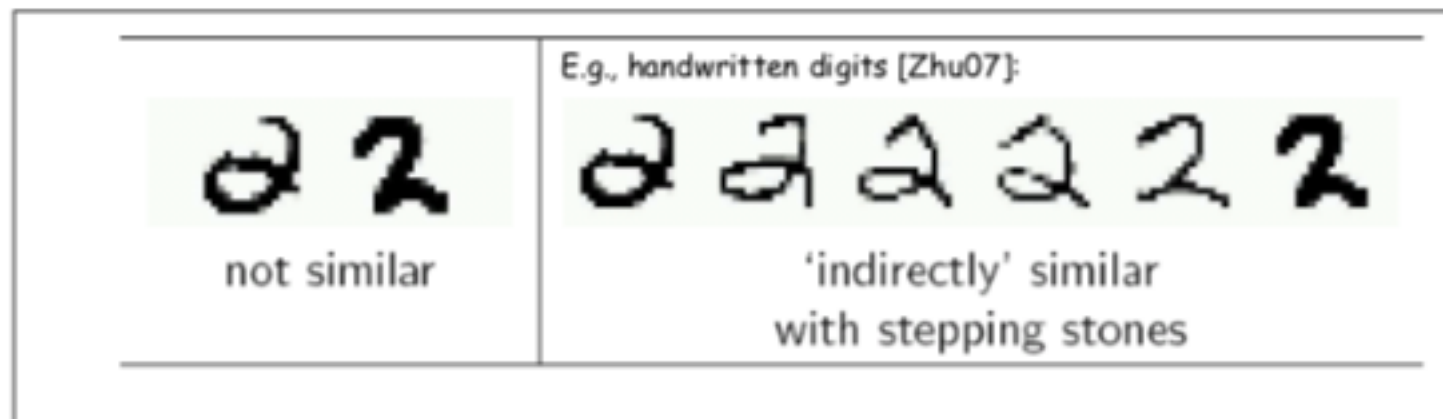
(b)



(c)

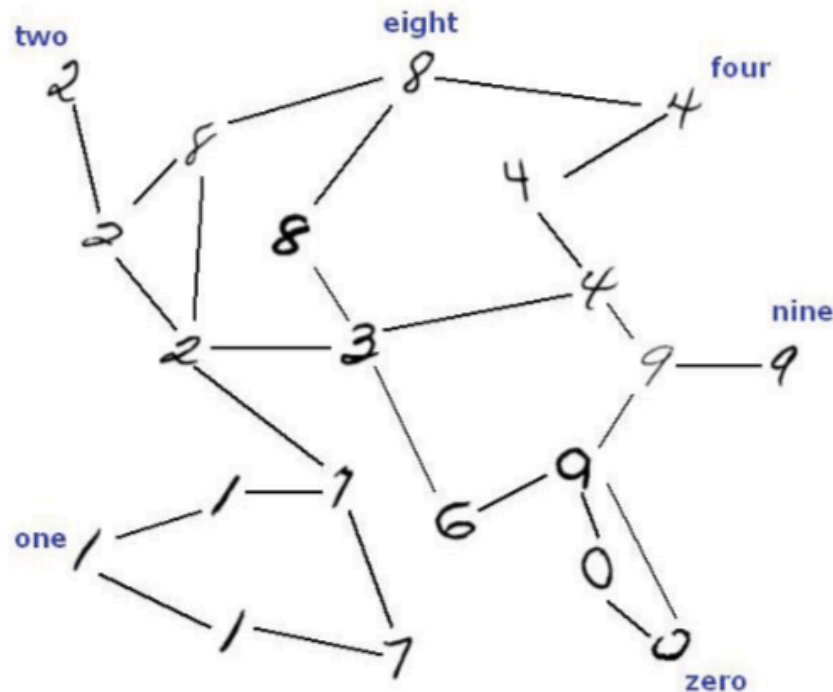
Graph based approaches

- Założenie podobieństwa przykładów ze względu na różne cechy – prawdopodobieństwo, że mogą mieć te same etykiety klas
- Dla wystarczająco dużo przykładów etykietowanych – podejścia NN sąsiedztwa
- Kiedy dodatkowo dużo nieetykietowanych – pomysł na stopniowe “przejście” tzw. Stepping stones examples
- Przykład rozpoznawania liter [Zhu]



Graph-based approaches

- Skonstruować graf, gdzie krawędzie łączą najbardziej podobne (wg. cech) przykłady
- Nieetykietowane przykłady łączą przykłady potencjalnie z tej samej klasy



Graph-based approaches

Unlabeled data can help “glue” the objects of the same class together.

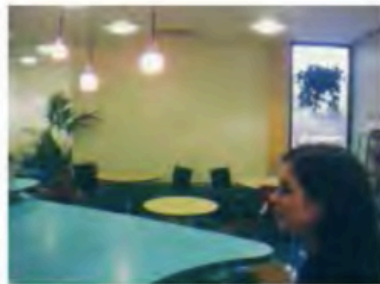
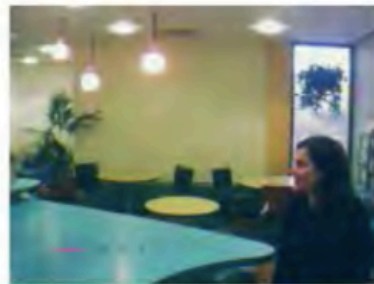


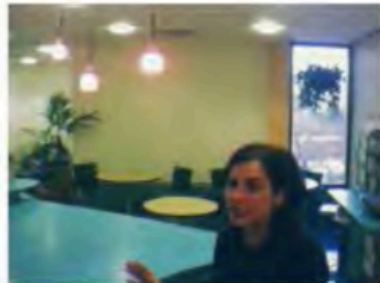
image 4005



neighbor 1: time edge



neighbor 2: color edge



neighbor 3: color edge



neighbor 4: color edge

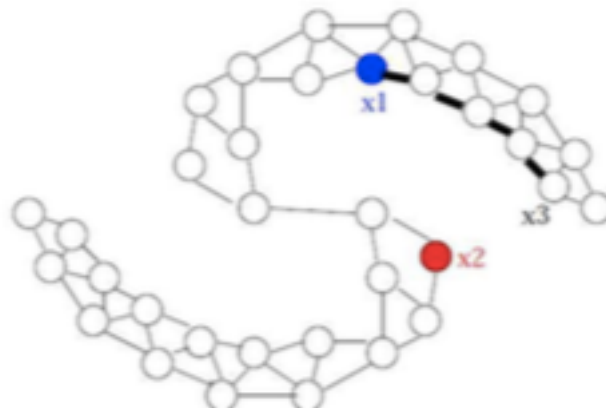


neighbor 5: face edge

Za: Person Identification in Webcam Images: An Application of Semi-Supervised Learning. [Balcan,Blum,Choi, Lafferty, Pantano, Rwebangira, Xiaojin Zhu], ICML 2005

Graf – sąsiedztwo przykładów

- Węzły $X_l + X_u$
- Podobieństwo – obliczane na podstawie cech
 - Graf k-nn, krawędzie najczęściej nieważone
 - Grał pełniejszy, redukcja wg. odległości, ważne krawędzie wg. Kernela gausowskiego
- Hipoteza: krawędzie o dużej wadze, przykłady o tych samych etykietach



Pomysły na konstrukcje grafu

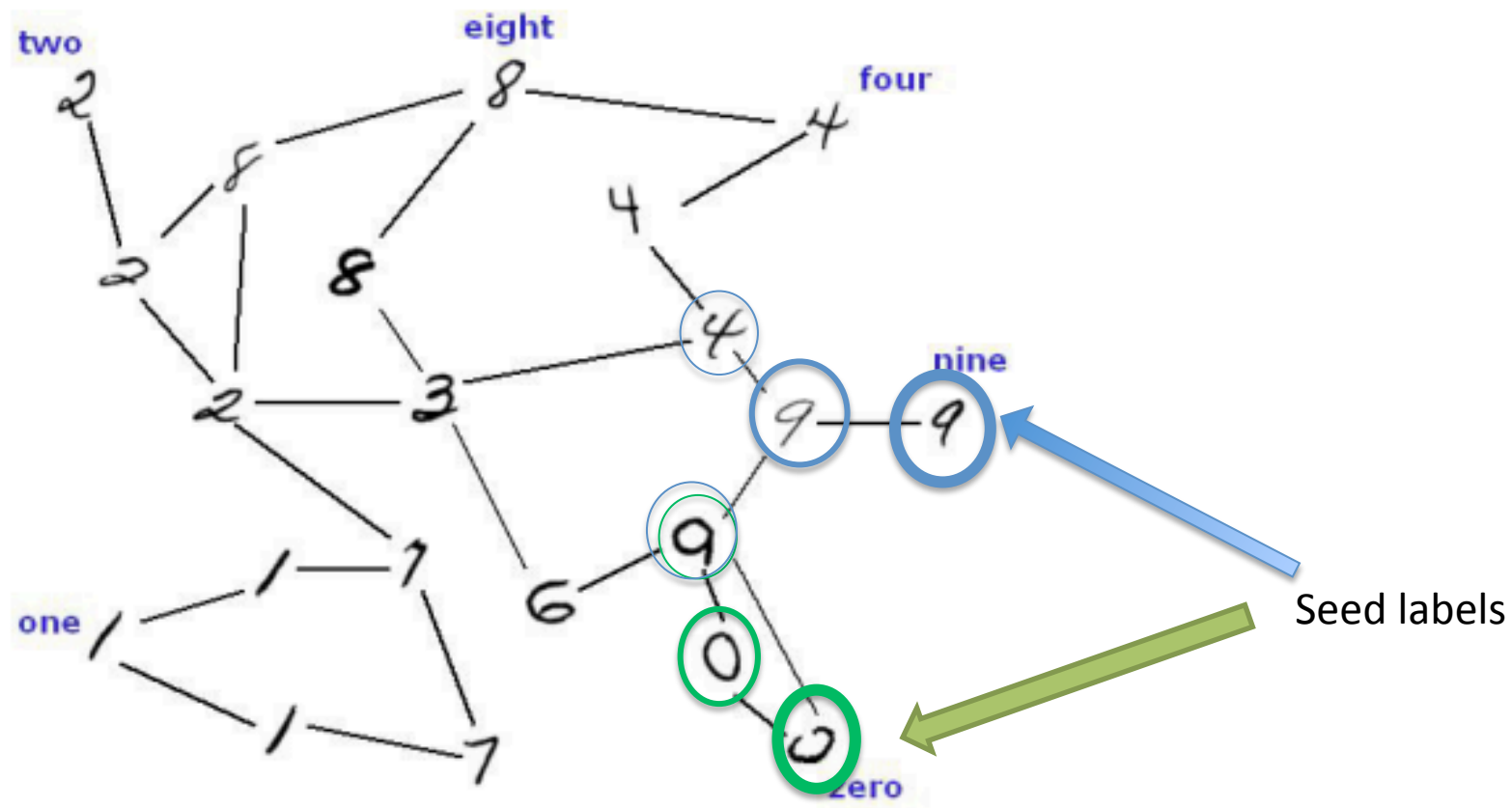
Empirically, the following works well:

1. Compute distance between i, j
2. For each i , connect to its k NN. k very small but still connects the graph
3. Optionally put weights on (only) those edges

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

4. Tune σ

SSL via “Label Propagation”

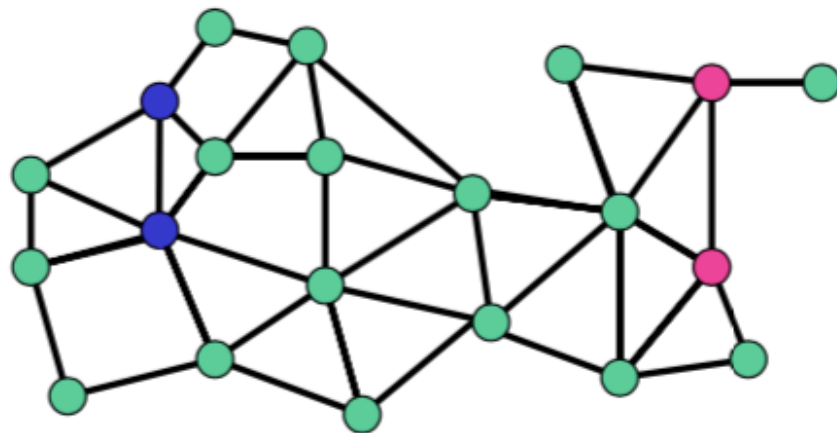


Graph based labeling

Graf z krawędziami pomiędzy najbliższymi przykładami
Różne podejścia do etykietowania węzłów

- Minimal cut
- Soft-cut $\sum_{e=(u,v)} (f(u) - f(v))^2$
- Special graph partitioning
- Regularyzacja i optymalizacja

Penalizacja modeli f , w których
nie zgadzają się etykiety
najbliższych sąsiadów



Przykład podejścia

- Assume the predictions on the entire data $\mathcal{L} \cup \mathcal{U}$ to be defined by function f
- Graph regularization assumes that **the function f is smooth**
 $\Rightarrow$ Similar examples i and j should have similar predictions f_i and f_j
- Graph regularization optimizes the following objective:

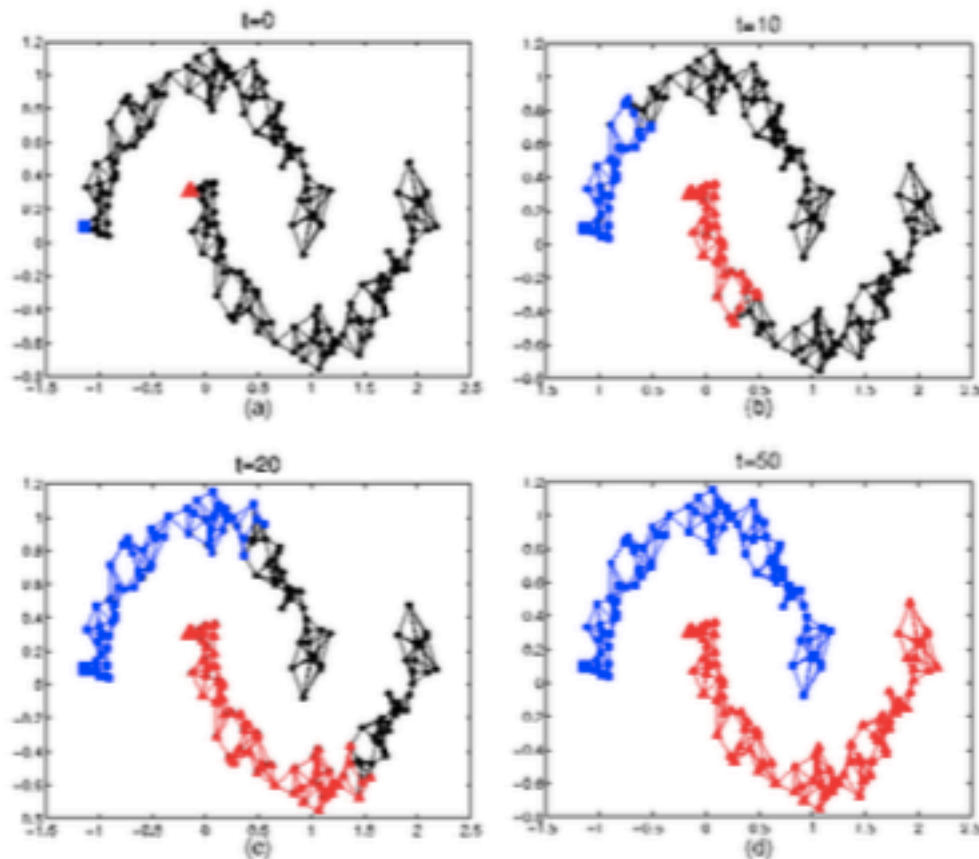
$$\min_f \sum_{i \in \mathcal{L}} (y_i - f_i)^2 + \lambda \sum_{i,j \in \mathcal{L}, \mathcal{U}} w_{ij} (f_i - f_j)^2$$

- First part is **minimizing the loss on labeled data**, second part **ensures smoothness of labels of labeled and unlabeled data**
 $\Rightarrow$ Minimization makes f_i and f_j to be very similar if w_{ij} is large
- λ is a trade-off parameter
- Several variants and ways to solve the above problem (refer to the SSL survey paper's section on graph based methods)

Więcej w cytowanych wykładach

Label propagation

Called **Label Propagation**, as the same solution is achieved by iteratively propagating labels along edges until convergence



Note: here
color $\hat{=}$ classes

[images from "Label Propagation Through Linear Neighborhoods", Wang, Zhang, ICML 2006]

Uczenie się w tzw. manifold

- Manifold – tzw. rozmaiłość n-wymiarowa
- Przejście z danych wysoce-wielowymiarowych w małowymiarową i skorzystanie z podobieństwa punktów

Manifold Assumption

1. The data lie on (or close to) a low-dimensional manifold.
2. Its intrinsic distance is relevant for classification.



[images from "The Geometric Basis of Semi-Supervised Learning", Sindhwani, Belkin, Niyogi
in "Semi-Supervised Learning" Chapelle, Schölkopf, Zien]

Algorithmic idea: use **Nearest-Neighbor Graph**

Uwagi nt. Oprogramowania

spójrz też na

<https://awesomeopensource.com/projects/semi-supervised-learning>

[Install](#) [User Guide](#) [API](#) [Examples](#) [More ▾](#)

[Prev](#) [Up](#) [Next](#)

scikit-learn 0.22.2
[Other versions](#)

Please [cite us](#) if you use the software.

1.14. Semi-Supervised
1.14.1. Label Propagation

1.14. Semi-Supervised

Semi-supervised learning is a situation in which in your training data some of the samples are not labeled. The semi-supervised estimators in `sklearn.semi_supervised` are able to make use of this additional unlabeled data to better capture the shape of the underlying data distribution and generalize better to new samples. These algorithms can perform well when we have a very small amount of labeled points and a large amount of unlabeled points.

Unlabeled entries in `y`

It is important to assign an identifier to unlabeled points along with the labeled data when training the model with the `fit` method. The identifier that this implementation uses is the integer value `-1`.

1.14.1. Label Propagation

Label propagation denotes a few variations of semi-supervised graph inference algorithms.

A few features available in this model:

- Can be used for classification and regression tasks
- Kernel methods to project data into alternate dimensional spaces

`scikit-learn` provides two label propagation models: **LabelPropagation** and **LabelSpreading**. Both work by constructing a similarity graph over all items in the input dataset.

Więcej – artykuły przeglądowe

- J.Engelen, H.Hoos: A survey on semi-supervised learning. Machine learning 2020.
- Zha et al.: Graph-based semi-supervised learning with multiple labels (2009)
- Liu et al.: Robust and Scalable Graph-Based Semisupervised Learning (2012)