Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

# Graph Construction for Semi-Supervised Learning

**Lilian Berton and Alneu de Andrade Lopes**
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
13560-970 São Carlos, SP - Brazil
Email: {lberton, alneu}@icmc.usp.br

## Abstract

Semi-Supervised Learning (SSL) techniques have become very relevant since they require a small set of labeled data. In this scenario, graph-based SSL algorithms provide a powerful framework for modeling manifold structures in high-dimensional spaces and are effective for the propagation of the few initial labels present in training data through the graph. An important step in graph-based SSL methods is the conversion of tabular data into a weighted graph. The graph construction has a key role in the quality of the classification in graph-based methods. Nevertheless, most of the SSL literature focuses on developing label inference algorithms without studying graph construction methods and its effect on the base algorithm performance. This PhD project aims to study this issue and proposes new methods for graph construction from flat data and improves the performance of the graph-based algorithms.

## 1 Introduction

Neighborhood graphs have been used in many areas to model local relationships for flat data. Usually, two approaches appear in the literature for constructing similarity based graphs: $\epsilon$-neighborhood and $k$-Nearest Neighbors (kNN) [Chapelle *et al.*, 2010]. An $\epsilon$-neighborhood graph is built connecting all the data points whose distance are smaller than $\epsilon$. These graphs are very sensitive to the parameter $\epsilon$ chosen and produce unusual degree distribution. The $k$NN graphs have better properties but still will always connect $k$ neighbors regardless of whether they are in the space. There are also the mutual $k$NN (M$k$NN) in which there is a connection between two vertices only if the rule of nearest neighbor is reciprocal. Hence, the mutual $k$NN is considered more restrictive and it is traditionally used in unsupervised learning.

The purpose of this PhD project is to extend the exploration of graph-based SSL algorithms, developing new techniques for graph construction from flat data. We are looking for answers for these questions discussed in the area: "Which graphs do we want to use to model our data? Which properties of graphs are attractive for ML? Which ones are misleading? Do algorithms behave differently on different kinds of graphs?". We tackle this issue into the following objectives: (i) Review graph construction strategies from the literature; (ii) Explore previous information available in the SSL for the graph construction; (iii) Investigate the role of the vertices degree in the network formation and their influence in the label inference algorithms; (iv) Investigate the relationship of networks generated with semi-supervised algorithms; (v) Apply the methods proposed in databases with a large number of instances and dimensions.

## 2 Proposed approaches

Following are our proposals for graph construction for SSL.

### 2.1 Regular graph construction

As $k$NN method greedily connects the $k$ nearest neighbors to each vertex and may return graphs where some vertices have more than $k$ neighbors, $b$-matching was proposed by [Jebara *et al.*, 2009], which ensures the graph is regular (every vertex with $b$ neighbors) and by experimental results the authors suggest that a regular graph can achieve better classification results compared to $k$NN. The authors cited an implementation whose guaranteed running time is $O(bn^3)$. In some cases, like in the work of [Ozaki *et al.*, 2011], building a $b$-matching graph is impracticable in terms of computational cost. We tackle this problem introducing an alternative method for generating regular graphs with better runtime performance $O(n^2)$ [Vega-Oliveros *et al.*, 2014]. Our technique is based on the preferential selection of vertices according some topological measures, like closeness, generating at the end of the process a regular graph. Experiments show that our method provides better or equal classification rate in comparison with $k$NN. Further, we employed the $k$NN, M$k$NN and the proposed method for regular graphs (S-$k$NN) in the relational algorithms for music genre classification [Valverde-Rebaza *et al.*, 2014]. Relational representations explore information about the instances that go beyond the attribute values, as they operate on graph models from the data. In our experiments, relational classifiers outperformed traditional classification techniques and the proposed method for graph construction leads to better classification accuracy.

### 2.2 Graph construction based on labeled instances

Most of the graph construction methods for SSL are unsupervised, i.e. they do not employ available label information during the graph construction process. Labeled data may be seen as a type of prior information which can be useful for improving graph construction for the current learning task. We proposed a method for graph construction that uses the available labeled data [Berton and de Andrade Lopes, 2014], denominated *Graph-based on informativeness of labeled instances* (GBILI). The proposed technique (GBILI) leads to good classification accuracy and has a quadratic time complexity. A parameter sensitivity analysis varying $k$ from 1 to 50 shows that for $k > 10$ GBILI presents stability in classification accuracy. Set parameters is a problem for many methods, GBILI has an advantage in this point. Analyses about network density show that $k$NN graphs become very dense as the value of $k$ increases. In contrast, GBILI graph converges for a constant average degree (around 2) despite the value of $k$. Besides, by using the $\phi$-edge ratio measure [Ozaki *et al.*, 2011], when the parameter $k$ becomes higher, the number of edges connecting vertices with different labels increases in $k$NN graphs, resulting in propagation of wrong label information. This situation does not happen in GBILI graphs. GBILI method leads the labeled points to become hubs. It is indicated by calculating centrality measures, like *node degree*, *betweenness*, *eigenvector* and *pageRank*. These measures are related with diffusion processes in a network, like information or disease spreading. As the labeled points in GBILI graphs are hubs they facilitate the label propagation.

### 2.3 Graph construction via link prediction

An important scientific issue regarding network analysis that has attracted attention in recent years is the link prediction. This problem aims to estimate the likelihood of the future existence of a link between two disconnected vertices. Considering the fact that link prediction is a mechanism for analyzing the growth and quick changes over time in underlying structures of the networks, it is feasible to think that link prediction can be used as a framework to evolve an initial neighborhood graphs constructed from tabular data. Hence, we propose a novel method for graph construction based on link prediction [Berton *et al.*, 2015]. First, an initial graph structure over the flat data set is built using some traditional graph construction technique. After, some link prediction method is computed with the objective of estimate new links in the graph. We show as our proposal improves the quality of graphs leading to better classification accuracy in supervised and semi-supervised domains.

## 3 Results

We compare the classification results using traditional methods for graph construction such as: $k$NN, M$k$NN, Minimum spanning tree (MST), $b$-matching ($b$M) and the proposed methods based on: link prediction ($k$NN+LP, MST+LP), regular graphs (S$k$NN) and labeled vertices (GBILI). Local and Global Consistency was used for label propagation task. The results considering the datasets proposed by [Chapelle *et al.*, 2010] and 10 labeled data are shown in Table 1. When the proposed methods achieve better accuracy than the best literature method, the results are in bold.

Table 1: SSL classification results

| **Method** | **g241c** | **g241n** | **Digit$_1$** | **USPS** |
|---|---|---|---|---|
| $k$NN | $0.544 \pm 0.06$ | $0.52 \pm 0.03$ | $0.894 \pm 0.05$ | $0.838 \pm 0.03$ |
| M$k$NN | $0.512 \pm 0.03$ | $0.515 \pm 0.02$ | $0.89 \pm 0.02$ | $0.841 \pm 0.06$ |
| MST | $0.499 \pm 0.01$ | $0.499 \pm 0.01$ | $0.499 \pm 0.01$ | $0.709 \pm 0.02$ |
| $b$M | $0.553 \pm 0.04$ | $0.534 \pm 0.04$ | $0.812 \pm 0.11$ | $0.823 \pm 0.03$ |
| $k$NN+LP | $\mathbf{0.581 \pm 0.07}$ | $0.524 \pm 0.03$ | $\mathbf{0.899 \pm 0.06}$ | $\mathbf{0.843 \pm 0.03}$ |
| MST+LP | $0.535 \pm 0.04$ | $0.51 \pm 0.16$ | $\mathbf{0.917 \pm 0.04}$ | $\mathbf{0.845 \pm 0.06}$ |
| S$k$NN | $\mathbf{0.572 \pm 0.06}$ | $0.53 \pm 0.03$ | $0.843 \pm 0.09$ | $0.822 \pm 0.06$ |
| GBILI | $\mathbf{0.569 \pm 0.08}$ | $\mathbf{0.568 \pm 0.08}$ | $0.843 \pm 0.06$ | $\mathbf{0.85 \pm 0.05}$ |

## 4 Final remarks

Many graph-based methods for SSL have been proposed, however studies involving the influence of the graph construction in such algorithms have received little attention. This PhD project investigated these aspects and proposed new techniques for graph construction exploring different characteristics, such as the influence of vertices degree, the available labeled vertices, graph measures and correlation among vertices. Experimental results performed indicate these proposals are promising and deserve further investigation.

## Acknowledgments

Grant 2011/21880-3 Sao Paulo Research Foundation (FAPESP).

## References

[Berton and de Andrade Lopes, 2014] Lilian Berton and Alneu de Andrade Lopes. Graph construction based on labeled instances for semi-supervised learning. In *22nd International Conference on Pattern Recognition*, pages 2477–2482, 2014.

[Berton *et al.*, 2015] Lilian Berton, Jorge Valverde-Rebaza, and Alneu de Andrade Lopes. Link prediction in graph construction for supervised and semi-supervised learning. In *International Joint Conference on Neural Networks*, page to appear, 2015.

[Chapelle *et al.*, 2010] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[Jebara *et al.*, 2009] Tony Jebara, Jun Wang, and Shih-Fu Chang. Graph construction and b-matching for semi-supervised learning. In *26th Annual International Conference on Machine Learning*, pages 441–448, 2009.

[Ozaki *et al.*, 2011] Kohei Ozaki, Masashi Shimbo, Mamoru Komachi, and Yuji Matsumoto. Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data. In *International Conference on Computational Natural Language Learning*, pages 154–162, 2011.

[Valverde-Rebaza *et al.*, 2014] Jorge Valverde-Rebaza, Aurea Soriano, Lilian Berton, Maria Cristina Ferreira de Oliveira, and Alneu de Andrade Lopes. Music genre classification using traditional and relational approaches. In *Brazilian Conference on Intelligent Systems*, pages 259–264, 2014.

[Vega-Oliveros *et al.*, 2014] Didier Augusto Vega-Oliveros, Lilian Berton, Andre Eberle, Alneu de Andrade Lopes, and Liang Zhao. Regular graph construction for semi-supervised learning. In *Journal of Physics: Conference Series*, volume 490, pages 012022–1–012022–4, 2014.