Machine learning via dynamical processes in complex networks*

Thiago Henrique Cupertino[1], **Zhao Liang**[2]

[1]Instituto de Ciências Matemáticas e Computação (ICMC)
Universidade de São Paulo (USP)
CEP 13566-590 – São Carlos – SP – Brazil

[2]Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP)
Universidade de São Paulo (USP)
CEP 14040-900 – Ribeirão Preto – SP – Brazil

thcupertino@gmail.com, zhao@usp.br

Abstract. *Machine learning is a research field devoted to the development of techniques capable of enabling a machine to "learn" from data. Many techniques have been proposed so far, but there are still issues to be unveiled specially in interdisciplinary research. In the doctorate thesis summarized here, we explore the advantages of network data representation to develop learning techniques based on dynamical processes. Our studies covered the three main learning categories: supervised, semi-supervised and unsupervised. We also applied the developed techniques for images, handwritten digits and letters recognition, and many other classification tasks. Advantages of network data representation and dynamical processes for machine learning are highlighted in all cases.*

1. Introduction

The main goal of the doctorate thesis summarized here was to develop new machine learning techniques by exploring advantages of dynamical processes on complex networks. Therefore, three main research areas were involved: **machine learning**, **complex networks** and **dynamical processes**. Also, the studies covered the three main categories of machine learning: supervised, semi-supervised and unsupervised methods.

In the machine learning domain, the topological structure is quite useful to detect diverse cluster or class shapes via a data clustering or classification algorithm, respectively. As a consequence, the application of network-based methods in learning tasks has been increasing over the past years and has become a active research area with a myriad of applications such as supervised learning [Bertini et al. 2011], semi-supervised learning [Breve et al. 2012], clustering [Silva et al. 2013] and dimensionality reduction [Yan et al. 2007].

For instance, traditional clustering techniques are very efficient in some cases, although they may be not suitable for some others. As an example where such algorithms are not appropriate is the clustering of data sets that contains groups of different spatial shapes, sizes and densities, as those methods generally assume that the clusters are

*Doctoral dissertation approved at the Instituto de Ciências Matemáticas e Computação - Universidade de São Paulo, for the degree of the Doctorate Program in Computer Science and Computational Mathematics, in December, 2013. Thesis site (click here).

hyper-spherical and have similar sizes. On the other hand, network-based data clustering methods are able to detect groups with arbitrary shapes because networks are powerful tools to represent topological relations among objects.

The motivations for researching supervised network-based techniques is mainly due to the possibility of different classification heuristics. Traditional classification techniques divide the data space according to physical features of the training data (similarity, distance, or distribution), that is, they divide the data space into subspaces, each one representing a data class. Strong twisting or largely overlapped subspaces are not permitted. In this way, many intrinsic and semantic relations among data items are ignored as, for example, topological structures. On the other hand, a relevant advantage when using networks is that it can perform quite different classification heuristics, taking into account topological structures, for example. Furthermore, the number of network-based supervised techniques reported so far is still small [Bertini et al. 2011].

In summary, network-based techniques presents some salient advantages over other traditional techniques for all the three main categories of machine learning. In the reported thesis, we also showed how the developed techniques can be used in various applications such as image, handwritten digits and letters recognition.

2. Main results

During the doctorate period, the investigation efforts were threefold: to explore and formulate applications for supervised, semi-supervised and unsupervised learning paradigms. Specifically, the achieved results were as follows:

Results for network-based supervised learning:

1. A new supervised classification technique which takes into account the **ease of access** of unlabeled instances to training classes through an underlying network. Differently from traditional classification heuristics, the proposed scheme uses random walk limiting probabilities to measure the limiting state transitions among training nodes;
2. A new supervised classification technique which combines the **ease of access** heuristic and the network **topological structure** to characterize data classes.
3. Network-based techniques were applied to the classification of **multiple observation sets**. For example, a car being tracked via a road camera system or a person having his/her images stored by an internal vision sensors network are examples of objects captured at different time instants or at different angles and geometric transformations, composing multiple observation sets.
4. Study of the usage of different network formation methods into a graph embedding framework to perform supervised **dimensionality reduction**. Some kinds of data, specially images, are often high-dimensional patterns. Dimensionality reduction can enhance processing and also increase classification accuracy.

Results for network-based semi-supervised learning:

1. The previous **ease of access** heuristic was extended to the semi-supervised case, in which each label is propagated through a network of unlabeled instances via a

biased random walk. The random walk process measures the label propagation from labeled vertices to the remainder unlabeled vertices of the network.

2. A new nature-inspired semi-supervised classification technique based on **attraction forces**. Labeled instances are considered as fixed attraction points that apply attraction forces on unlabeled instances. In turn, the unlabeled instances are expected to move towards the resultant force direction and, eventually, to converge to an attraction point.

Results for network-based unsupervised learning:

1. A new data clustering technique based on **synchronization and pinning control** of networked dynamical oscillators. A dissimilarity measure is computed via a dynamic system in which vertices are expected to reach a consensus state regarding a reference trajectory forced into the system via pinning control. The resultant dissimilarity provides a set of features for the data items allowing the detection of clusters with different shapes and sizes.

Henceforward, we describe briefly only the main results concerning the developments for supervised learning due to space restriction. The complete doctorate thesis can be found at thesis site (click here).

We introduced a new network-based classification technique which considers the ease of access heuristic and network structure measurements to perform **high-level classification**. The proposed technique uses a measure for the dynamical process called random walk limiting probabilities through an underlying network that combines both data similarity and structural measures.

The proposed approach is actually a general classification scheme in such a way that one can put any new classification criteria in the weight matrix to guide the random walk process. We showed how the low level classification term, represented by similarities between data items, and the high level term, represented by a network structural measure, are combined in this scheme. As a consequence, the proposed approach can classify data not only using physical features, but also checking the structural pattern formation via the network constructed from the training data. We also showed that such a combination can improve the classification results in real applications, such as image recognition. Another interesting finding of this work is that the high level term embedded in the connection matrix of the data network is specially useful in complex situations where the low level term fails such as when there is a high mixture of data classes.

For example, the proposed technique is able to classify the black squares, which complement the straight line started by the green discs in Fig. 1. This example depicts a situation in which the **network structural term** plays a key role in classification.

The efficacy of the technique is also verified on each category of the ETH-80 data set depicted in Fig. 2. As each category comprises very similar classes, the classification process is a hard task. The results showed that a combination of the similarity and the structural weight matrices resulted in the best classification accuracies.

3. Conclusions

The summarized doctoral thesis have joined three main research areas: machine learning, complex networks and dynamical processes. The main goal has been to enhance and de-

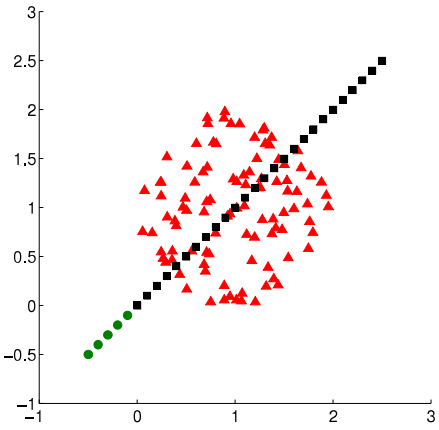

Figure 1. Artificial data set composed of two overlapped classes forming specific patterns. The objective is to classify the black squares as belonging to the straight line including the green discs class, which can be performed by the proposed technique.

Figure 2. ETH-80 images collection. Each line corresponds to a category of objects. Each category comprises 10 distinct classes.

velop machine learning techniques by exploring the advantages of network representation and dynamical processes on networks. The motivations came from the fact that network representation unifies the structure, dynamics and functions of the system it represents, and it is able to evidence topological structures. These characteristics make networks suitable for machine learning tasks by revealing intrinsic structures and their evolutions within data set relationships. The proposed methods differ from traditional classification in which decision borders are defined to separate groups in the attribute space. Furthermore, the data are not classified by using only its physical attributes, but also using the semantic structure formed by the data in a network. Hence, even if different classes are overlapped in the attribute space, they can be correctly identified by taking into account the semantic structure each class comprises.

In this report, only a part of the supervised learning developments have been introduced due to space restriction. For a complete explanation concerning the other developments, including the semi-supervised and unsupervised techniques and applications,

please refer to the thesis site (click here).

4. List of publications

The investigations reported in the thesis have been published in scientific articles in many international journals and conferences, totaling 16 articles. A grouped list by subject is given below:

Articles for supervised learning techniques

- Articles in international journals:
    1. Cupertino, T. H. and Zhao, L. A unified scheme for data classification using random walk in networks. IEEE Transactions on Cybernetics. Submitted, 2014.
    2. Cupertino, T. H. and Zhao, L. Network-based supervised data classification by using an heuristic of ease of access. Neurocomputing. Accepted for publication, 2014.
    3. Cupertino, T. H.; Silva, T. C.; Zhao, L. Classification of multiple observation sets via network modularity. Neural Computing & Applications (Internet), p. 1-7, 2012. DOI: 10.1007/s00521-012-1115-y.
- Publications in international conference proceedings:
    1. Carneiro, M. G.; Cupertino, T. H.; Zhao, L. k-Associated optimal network for graph embedding dimensionality reduction. IEEE World Congress on Computational Intelligence (WCCI), July 2014, Beijing, China. Accepted.
    2. Cupertino, T. H.; Carneiro, M. G.; Zhao, L. Dimensionality reduction with the k-associated optimal graph applied to image classification. IEEE International Conference on Imaging Systems and Techniques (IST), October 2013, Beijing, China.
    3. Cupertino, T. H.; Zhao, L. Bias-guided random walk for network-based data classification. In: Tenth International Symposium on Neural Networks (ISNN), 2013, Dalian, China, v. 7952, p. 375-384. DOI: 10.1007/978-3-642-39068-5_46.
    4. Cupertino, T. H.; Zhao, L. Using Katz centrality to classify multiple pattern transformations. In: 2012 Brazilian Symposium on Neural Networks, Curitiba - PR, v. 1, p. 232-237, DOI: 10.1109/SBRN.2012.23.
    5. Cupertino, T. H.; Silva, T. C.; Zhao, L. Multiple images set classification via network modularity. In: SIBGRAPI 2012 - Conference on Graphics, Patterns and Images - Workshop of Theses and Dissertations, Ouro Preto - MG, v. 1, p. 124-129.
    6. Silva, T. C.; Cupertino, T. H. High level classification for pattern recognition. In: XXIV Sibgrapi Conference on Graphics, Patterns and Images, 2011, Maceió - AL. Los Alamitos: IEEE Computer Society, 2011. v. 1. p. 344-351. DOI: 10.1109/SIBGRAPI.2011.19.

Articles for semi-supervised learning techniques

- Articles in international journals:

1. Cupertino, T. H.; Gueleri, R.; Zhao, L. A semi-supervised classification technique based on interacting forces. Neurocomputing (Amsterdam), p. 1-9, 2013. DOI: 10.1016/j.neucom.2013.05.050.

- Publications in international conference proceedings:
  1. Gueleri, R.; Cupertino, T. H.; Carvalho, A. P.; Zhao, L. A flocking-like technique to perform semi-supervised learning. IEEE World Congress on Computational Intelligence (WCCI), July 2014, Beijing, China. Accepted.
  2. Cupertino, T. H.; Zhao, L. Semi-supervised learning using random walk limiting probabilities. In: Tenth International Symposium on Neural Networks (ISNN), 2013, Dalian, China, v. 7952, p. 395-404. DOI: 10.1007/978-3-642-39068-5_48.
  3. Cupertino, T. H.; Zhao, L. Using interacting forces to perform semi-supervised learning. In: 2012 Brazilian Symposium on Neural Networks, Curitiba - PR, v. 1, p. 91-96, DOI: 10.1109/SBRN.2012.24.

Articles for unsupervised learning techniques

- Articles in international journals:
  1. Cupertino, T. H.; Huertas, J., Zhao, L. Data clustering using controlled consensus in complex networks. Neurocomputing (Amsterdam), v. 118, p. 132-140, 2013. DOI: 10.1016/j.neucom.2013.02.026.
  2. Silva, T. C.; Zhao, L.; Cupertino, T. H. Handwritten data clustering using agents competition in networks. Journal of Mathematical Imaging and Vision, v. 45, n. 3, p. 264-276, 2012. DOI: 10.1007/s10851-012-0353-z.
- Publications in international conference proceedings:
  1. Silva, T. C.; Cupertino, T. H. Stochastic competitive learning applied to handwritten digit and letter clustering. In: XXIV Sibgrapi Conference on Graphics, Patterns and Images, 2011, Maceió - AL. Los Alamitos: IEEE Computer Society, 2011. v. 1. p. 313-320. DOI: 10.1109/SIBGRAPI.2011.35.

References

[Bertini et al. 2011] Bertini, J. R., Zhao, L., Motta, R., and Lopes, A. d. A. (2011). A non-parametric classification method based on k-associated graphs. *Information Sciences*, 181:5435–5456.

[Breve et al. 2012] Breve, F., Zhao, L., Quiles, M. G., Pedrycz, W., and Liu, J. (2012). Particle competition and cooperation in networks for semi-supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 24(9):1686–1698.

[Silva et al. 2013] Silva, T. C., Zhao, L., and Cupertino, T. H. (2013). Handwritten data clustering using agents competition in networks. *Journal of Mathematical Imaging and Vision*, 45(3):264–276.

[Yan et al. 2007] Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., and Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):40–51.