Data Classification in Complex Networks via Pattern Conformation, Data Importance and Structural Optimization

Murillo G. Carneiro[1,2], **Liang Zhao**[1,3]

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

[2]Faculdade de Computação, Universidade Federal de Uberlândia (UFU)
38400-902 – Uberlândia – MG – Brazil

[3]Departamento de Computação e Matemática, Universidade de São Paulo (USP)
14040-901 – Ribeirão Preto – SP – Brazil

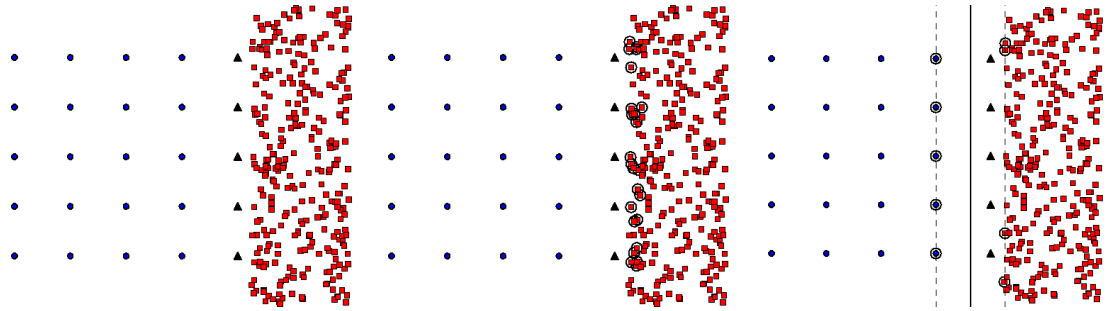
mgcarneiro@ufu.br, zhao@usp.br

Abstract. *Most data classification techniques rely only on the physical features of the data (e.g., similarity, distance or distribution), which makes them difficult to detect intrinsic and semantic relations among data items, such as the pattern formation, for instance. In this thesis, it is proposed classification methods based on complex networks in order to consider not only physical features but also capture structural and dynamical properties of the data through the network representation. The proposed methods comprise concepts of pattern conformation, data importance and network structural optimization, which are related to complex networks theory, learning systems, and bioinspired optimization. Extensive experiments demonstrate the good performance of our methods when compared against representative state-of-the-art methods over a wide range of artificial and real data sets, including applications in domains such as heart disease diagnosis and semantic role labeling.*

1. Introduction

This paper summarizes the main contributions of the doctorate research presented in [Carneiro 2016], with complex networks and machine learning as major topics. By the ubiquitous nature and by providing a set of efficient and robust tools to model and analyze networked data, complex networks have become a promising research topic for many areas, including machine learning and data mining [Silva and Zhao 2016]. Typical examples of network-based learning include unsupervised and semi-supervised tasks, such as community detection (or data clustering), label propagation and dimension reduction [Chapelle et al. 2006, Fortunato 2010]. Although data classification is a largely investigated task, the development of supervised learning methods based on complex networks is also a barely explored topic. In such a task, there is no space for label or other information propagation process into the network as there is only one or very few unlabeled data items [Carneiro 2016]. In addition to cover this lack, the following issues are also addressed by the investigations presented in the thesis:

Problem: The literature contains a myriad of data classification techniques. Traditionally, these techniques define decision boundaries in the data space according to the physical features of a training set and a new data item is classified by verifying its relative position



(a) Two-class data set (b) k-NN (**Nearest neighbors**) (c) SVM (**Support vectors**)

Figure 1. Analysis of the classification process of traditional techniques in a simple two class data set where circle data items denote a clear pattern and triangle data items need to be classified. Such techniques fail to consider the semantic structure of the data and they consequently label triangle data items as belonging to square/red class.

to the boundaries. Such kind of classification, which is only based on the physical attributes of the data (e.g., similarity, distance or distribution), makes traditional techniques unable to detect intrinsic and semantic relations among the data items such as the pattern formation, for instance. Let us consider Fig. 1(a), which shows a simple data set with two classes denoted by the circle and the square data items. The triangle data items represent test instances that need to be classified. Figs. 1(b) and 1(c) shows the classification process behind the traditional techniques k-nearest neighbors (k-NN) and Support Vector Machine (SVM). For example, k-NN classifies a test instance by verifying the label of its k nearest neighbors (circulated in Fig. 1(b)), and SVM takes into account the support vectors (circulated in Fig. 1(c)) to approximate each class with a convex hull and to find the best separating hyperplane between the classes. One can see both techniques fail to identify semantic patterns formed by the data. By contrast, the usage of complex networks is a promising way to capture spatial, topological and functional relationships of the data, as the network representation unifies structure, dynamic and functions of the networked system [Newman 2010].

Pattern conformation: In supervised data classification, the first attempt to use complex networks in order to consider the semantic relations among the data items is the hybrid framework for highlevel classification proposed in [Silva and Zhao 2012]. Such framework combines the associations produced by traditional and network-based techniques. The network-based technique uses complex network measures to estimate the membership of a test item according to the data formation pattern. However, given the high number of parameters in such framework (e.g., network formation, network measures variation, parameters of the traditional technique, convex combination, and so on), a simplified framework for highlevel classification which employs an unique network to provide both physical and complex-network based associations is proposed in the thesis .

Data importance: Although pattern conformation has been employed by highlevel classification as a new classification concept, other concepts can also be derived from complex networks. For example, structural and dynamical properties of the networked data can provide additional layers of information by defining quantitatively the importance of each data item into the classification process. Despite it has been a common practice in

data classification to assume that all data samples have the same relevance, such an assumption is obviously not compliant to the natural classification performed by the human brain. Moreover, neglecting the individual importance of each data sample may change the understanding of the whole data set. Such point is addressed in [Carneiro 2016] which proposes a new classification concept embedded in the constructed networks, considering both physical and topological features of the input data and permiting the detection of a variety of data patterns using the importance measure derived from Google's PageRank.

Network optimization: A common characteristic among network-based methods is the construction of the underlying data graph. In machine learning, such graph is usually formed from the input vector data, where each data item is represented as a node and the edges are defined from the affinity (or similarity) among the data items, for example, by connecting each data item to its k nearest neighbors. Although network formation is a crucial step for good performance, few attention has been devoted to this topic [Newman 2010]. Such a situation is confirmed by the common usage of the simple kNN network construction method in literature. kNN method considers only local data relationships and it is a general-purposed one, i.e., the constructed networks can be used for any machine learning or data mining tasks. By contrast, sophisticated methods consider both local and global relationship of the input data, but they are restricted for specific purposes. In order to fill this gap, an optimization framework, which is responsible to construct an "optimal" network regarding a given processing goal, is presented in [Carneiro 2016, Carneiro et al. 2016a].

In summary, this doctorate research investigated whether the structural and dynamical features derived from network representation can provide efficient computational methods to sort out the issues discussed above. The main contributions derived from this thesis are briefly discussed in the next section.

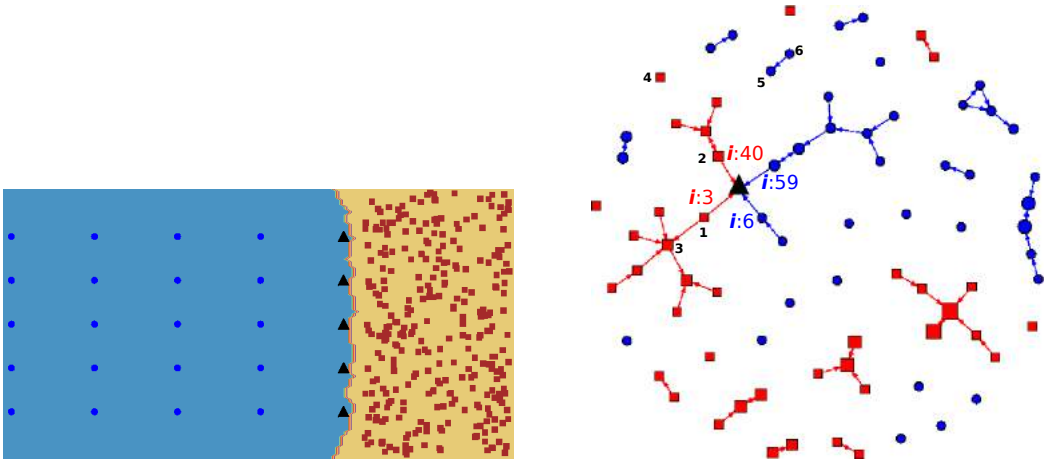
2. Thesis Contributions

Most of the doctorate research has been focused on data classification via importance concept in complex networks and network structural optimization. Following both contributions are briefly discussed. As extensive experiments were conducted for each proposed method, one may refer to [Carneiro 2016] in order to obtain a complete and precise description about the experimental setting and comparative analysis.

Classification Based on Data Importance. This investigation proposed a new data classification concept based on the importance concept of complex networks. Instead of data space division as having been done in traditional techniques or pattern conformation as having been done in highlevel classification techniques, the classification based on the data importance considers the individual importance of each data item in order to classify an unlabeled item. In the developed technique, the concept of importance is derived from PageRank, the ranking measure operating behind the universal search engine of Google. In addition, the technique captures spatial and structural properties of the networked data from a new network measure created, named spatio-structural differential efficiency. Briefly, in the training phase, the efficiency and PageRank measures are calculated over the underlying network constructed from the training data by using any graph construction method (e.g., kNN network). In the test phase, by using the spatio-structural differential efficiency measure, each test data item is temporarily connected to a set of

vertices in which its importance is then calculated for each class, and the test item is labeled into that class where it has the highest importance. The good performance of the proposed method is attested by comparisons against state-of-the-art methods over a wide range of artificial and real-world data sets, including applications in domains such as heart disease diagnosis [Carneiro 2016].

Fig. 2 gives a glimpse into the classification process of the proposed technique. Taking into account the artificial data set presented by Fig. 1(a), in which k-NN and SVM are unable to label the test data items correctly, Fig. 2(a) demonstrates that the importance-based method accurately detect the pattern formation of the data. A real-world example is also considered in Fig. 2(b), which illustrates the classification process of a patient (represented as △/black data item) in terms of heart disease diagnosis. In the figure, patients diagnosed with heart disease are represented by blue/○ markers; otherwise by red/□ markers. Despite the importance-based technique diagnoses the heart disease of the patient correctly, traditional techniques, such as k-NN, SVM, random forest, etc., fail in such task by considering only the physical features of the data. For example, in the same figure, the nearest neighbors of the new patient is showed. One can see k-NN classify the patient to the red/□ class, i.e., without heart disease. Thus, both examples in Fig. 2 show that the proposed technique contributes to the data classification task by considering the organizational structure of the data beyond the physical features. Moreover, the experimental results also revealed the low computational cost of the new technique in comparison to other traditional techniques widely used in literature.



(a) Classification of the data in Fig. 1(a)

(b) Heart disease diagnosis of a patient

Figure 2. Analysis of the classifications provided by the importance-based data classification over misleading cases for traditional techniques.

Bioinspired Network Optimization. In the thesis, it is also proposed a bioinspired optimization framework which is expected to build up the network while conducting the optimization of a task-oriented quality function. Such method can be divided in two phases. In the optimization phase (training), we designed a mapping heuristic to determine the whole set of possible network configurations which is employed by our framework to convert particles to networks and vice versa. The network configurations are then iteratively evaluated and updated by a particle swarm optimization method which returns the best particle, i.e., the best network configuration according to a given function. Two quality

functions are evaluated in the experiments: highlevel and importance-based classification. Results on artificial and real-world data sets (including semantic role labeling) reveal the network provided through the structural optimization presents statistically better results than those generated by the most used network formation methods in literature, especially in higher complexity of class configuration (such as the mixture among different classes). In addition, they also performed well in comparison with widely used traditional classification techniques, such as SVM and logistic regression [Carneiro et al. 2016a]. Moreover, the proposed framework can also be adapted to perform structural optimization for other graph-based learning tasks, such as dimension reduction and outlier detection.

Following we also list other relevant contributions discussed in the thesis:

- **A simplified framework for highlevel classification:** in the thesis, the highlevel classification is simplified in a proposed hybrid technique where physical and complex-network based associations are produced from the same network, reducing considerably the number of parameters [Carneiro et al. 2014b, Carneiro and Zhao 2013]. Experimental results show that a larger portion of the highlevel association is required to get correct classification when there is a complex-formed and well-defined pattern in the data set. They also demonstrate that the proposed technique presents competitive performance against state-of-the-art methods (e.g., SVM) and it outperforms typical data classification techniques (e.g., classification and regression trees).
- **Graph-based semantic role diffusion:** the scarcity of annotated data for Brazilian Portuguese semantic role labeling is taken into account in [Carneiro et al. 2016b], which proposed a graph-based semi-supervised framework based on label propagation in order to investigate the diffusion of semantic roles for that language; results show label propagation methods outperform a baseline.
- **Parameter-free graph-based dimension reduction (DR):** in [Carneiro et al. 2014a, Cupertino et al. 2013], it is proposed a parameter-free graph-embedding DR method whose results are competitive compared to classical network approaches (e.g., kNN) and widely used DR methods (e.g., principal component analysis).

3. Publications

Following it is provided a summary of the main articles originated during the doctorate research. It is also presented the Qualis of the journals and conferences according to the latest version released by CAPES*.

- Carneiro, M. G., Zhao, L., and Jin, Y. Bio-inspired structural optimization for network-based data classification (under review). *IEEE Trans. Cybern.* – **Qualis A1**.
- Carneiro, M. G., Zhao, L., and Rosa, J. L. G. Improving semantic role labeling using highlevel classification in complex networks (accepted). In *FSKD* – **Qualis B1**.
- Carneiro, M. G. and Zhao, L. Organizational data classification based on the importance concept of complex networks (under review). *IEEE Trans. Neural Netw. and Learn. Syst.* – **Qualis A1**.
- Carneiro, M. G., Zhao, L., Cheng, R., and Jin, Y. (2016a). Network structural optimization based on swarm intelligence for highlevel classification. In *IEEE IJCNN*, pages 3737–3744 – **Qualis A1**.
- Carneiro, M. G., Zhao, L., and Rosa, J. L. G. (2016b). Graph-based semi-supervised learning for semantic role diffusion. In *KDMiLe*, pages 108–115.
- Cupertino, T. H., Zhao, L., and Carneiro, M. G. (2015). Network-based supervised data classification by using an heuristic of ease of access. *Neurocomputing*, 149:86–92 – **Qualis A1**.

*Also available on `http://qualis.ic.ufmt.br/`

• Carneiro, M. G., Rosa, J. L. G., Lopes, A. A., and Zhao, L. (2014b). Network-based data classification: combining k-associated optimal graphs and high-level prediction. *J. Braz. Comp. Soc*, 20(1):1–14 – **Q. B1**.
• Carneiro, M. G., Cupertino, T. H., and Zhao, L. (2014a). K-associated optimal network for graph embedding dimensionality reduction. In *IEEE IJCNN*, pages 1660–1666 – **Qualis A1**.
• Cupertino, T. H., Carneiro, M. G., and Zhao, L. (2013). Dimensionality reduction with the k-associated optimal graph applied to image classification. In *IEEE IST*, pages 366–371 – **Qualis B2**.
• Carneiro, M. G. and Zhao, L. (2013). High level classification totally based on complex networks. In *IEEE BRICS-CCI*, pages 507–514.

Some other articles originated during the PhD period, which are not directly related to the thesis, are listed as follows.

• Carvalho, T. I., Carneiro, M. G., and Oliveira, G. M. B. Improving cellular automata scheduling through dynamic control (under review). *Int. Journal of Parallel, Emergent and Distributed Systems* – **Qualis B1**
• Carneiro, M. G. and Oliveira, G. M. B. (2013). Synchronous cellular automata-based scheduler initialized by heuristic and modeled by a pseudo-linear neighborhood. *Nat. Comput*, 12(3):339–351 - **Q. B1**

References

Carneiro, M. G. (2016). *Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural*. PhD thesis, Universidade de São Paulo.

Carneiro, M. G., Cupertino, T. H., and Zhao, L. (2014a). K-associated optimal network for graph embedding dimensionality reduction. In *IEEE IJCNN*, pages 1660–1666.

Carneiro, M. G., Rosa, J. L. G., Lopes, A. A., and Zhao, L. (2014b). Network-based data classification: combining k-associated optimal graphs and high-level prediction. *J. Braz. Comp. Soc*, 20(1):1–14.

Carneiro, M. G. and Zhao, L. (2013). High level classification totally based on complex networks. In *IEEE BRICS-CCI*, pages 507–514.

Carneiro, M. G., Zhao, L., Cheng, R., and Jin, Y. (2016a). Network structural optimization based on swarm intelligence for highlevel classification. In *IEEE IJCNN*, pages 3737–3744.

Carneiro, M. G., Zhao, L., and Rosa, J. L. G. (2016b). Graph-based semi-supervised learning for semantic role diffusion. In *KDMiLe*, pages 108–115.

Chapelle, O., Scholkopf, B., and Zien, A. (2006). *Semi-Supervised Learning*. MIT Press.

Cupertino, T. H., Carneiro, M. G., and Zhao, L. (2013). Dimensionality reduction with the k-associated optimal graph applied to image classification. In *IEEE IST*, pages 366–371.

Cupertino, T. H., Zhao, L., and Carneiro, M. G. (2015). Network-based supervised data classification by using an heuristic of ease of access. *Neurocomputing*, 149:86–92.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5):75 – 174.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc.

Silva, T. C. and Zhao, L. (2012). Network-based high level data classification. *IEEE Trans. Neural Netw. and Learn. Syst.*, 23(6):954–970.

Silva, T. C. and Zhao, L. (2016). *Machine Learning in Complex Networks*. Springer.