Information Sciences 294 (2015) 109–126

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

High-level pattern-based classification via tourist walks in networks

Thiago Christiano Silva *, Liang Zhao

Department of Computer Science and Mathematics, School of Philosophy, Science and Literature in Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo 14040-901, Brazil

ARTICLE INFO

Article history:
Received 6 March 2014
Received in revised form 19 August 2014
Accepted 17 September 2014
Available online 7 October 2014

Keywords:
High-level classification
Tourist walk
Supervised learning
Complex network
Machine learning
Pattern recognition

ABSTRACT

In this paper, we present a hybrid classification technique, which combines the decisions of low- and high-level classifiers. The low-level term realizes the classification task considering only the input data's physical features, such as geometrical or statistical characteristics. In contrast, the high-level classification process checks the compliance of the new test instances against the pattern formations of each class that composes the training data. For this end, we extract suitable organizational and topological descriptors of a network that is constructed from the input data. With these descriptors, we show that the high-level term has the ability of detecting data patterns with semantic and global meanings. Here, the input data's pattern formations are extracted by utilizing the dynamical information generated from several *tourist walk processes*, which are performed on the resulting network. Specifically, weighted combinations of transient and cycle lengths, which are derived variables from the tourist walks, are employed. Moreover, we show an effective method for calibrating the learning weights of these terms by using a statistical approach. Furthermore, we show that the tourist's memory size is related to what extent one may capture organizational and complex features of the network. This means that local, quasi-local, and global features can be extracted, depending on the value of memory size parameter. Still in this work, we uncover the existence of a critical memory length, here denominated *complex saturation*, where any values larger than this critical point make no changes in the behaviors of the transient and cycle lengths. We also investigate several artificial and real-world situations where the low-level term alone fails to identify intrinsic data patterns, but the high-level term is able to perform well. Our investigation suggests that the proposed technique is able to improve the already optimized performance of traditional classification techniques. Finally, we apply the proposed technique in recognizing handwritten digits images and interesting results are obtained.

1. Introduction

In supervised data classification, for a given training set, a map from the input data to the corresponding desired output is estimated. The constructed map, called a classifier, is used to predict new input instances. Many supervised data classification techniques have been developed [6,11,19,23,30,31,35,42,43], such as k-nearest neighbors, Bayesian classifiers, neural

* Corresponding author.
E-mail addresses: thiagochris@gmail.com (T.C. Silva), zhao@usp.br (L. Zhao).

http://dx.doi.org/10.1016/j.ins.2014.09.048

networks, decision trees, committee machines, spectral biclustering, genetic algorithms, gravitational-based methods, and so on. In essence, all these techniques train and, consequently, classify unlabeled data items according to the physical features (e.g., distance, similarity or distribution) of the input data. These techniques that predict class labels using only physical features are called *low-level classification* techniques [36].

Usually, data items are not isolated points in the attribute space, but instead tend to form certain patterns. For example, in Fig. 1, the two test instances represented by the triangles are most probably to be classified as members of the square-shaped class if only physical features, such as distances among data instances, are considered. On the other hand, if we take into account the relationships among the data, we would intuitively classify the triangle-shaped items as members of the circular-shaped class, since a clear pattern (lozenge) is formed. The human (animal) brain performs both low and high orders of learning and it has facility of identifying patterns according to the semantic meanings of the input data. In general, however, this kind of task is still hard to be performed by computers. Supervised data classification by considering not only physical attributes but also pattern formation is referred to as *high-level classification* [36].

Broadly speaking, low-level classification techniques often share the same heuristic: division of the data space into sub-spaces, each of which representing a class. They are short in reproducing complex-formed or twisted classes, because they often rely on assumptions such as fixed shapes or predefined distributions. In contrast, the salient feature of the proposed technique is that it really provides two distinct classification heuristics: low- and high-level classifications. The former performs the prediction by the data's physical features, while the latter captures the data's pattern formations, which, in turn, permits the classifier to reproduce complex-formed and (or) twisted classes. As a result, a test instance is declared as member of the class to which it complies in a structural sense, no matter how far it is from the center or any members of that class.

It is well known that the network representation can capture arbitrary levels of relationships or interactions of the input data [37–39]. For this reason, we here show how the networks' topological properties can help in identifying the pattern formation and, consequently, be used for general high-level classification. In this work, these topological properties are revealed by the *tourist walks*. A tourist walk can be defined as follows [21]. Given a set of cities, at each time step, the tourist (walker) goes to the nearest city that has not been visited in the past μ time steps. It has been shown that tourist walk is useful for data clustering [8] and image processing [3]. Each tourist walk can be decomposed in two terms: (i) the initial *transient part* of length t and (ii) a *cycle* (attractor) with period c. However, all these kinds of works are realized in regular lattices. Here, we study tourist walks in networks and we show that it has the ability of capturing the topological properties of the underlying network in a local to global fashion. It is worth observing that the application of tourist walks to graph-based environments is a new approach taken here. In addition, the employment of the tourist walks' dynamics for discovering patterns in networks is a totally novel scheme in the literature.

Following the literature stream on such matter, several kinds of works related to high-level classification may be highlighted, such as:

- the Semantic Web [4,12,34], which uses ontologies to describe the semantics of the data;
- statistical relational learning, which may be decomposed in methods that realize collective inference [15,25,45–47] or graph-based semisupervised learning [9,48];
- contextual classification techniques [5,10,24,27,40,41,44], which consider the spatial relationships between the individual pixels and the local and global configurations of neighboring pixels in an image for assigning classes.

All the above-mentioned techniques, on one hand, try to make inferences for a new data item in accordance with the neighborhood relationships between data samples (nodes) within the graph. On the other hand, our approach aims at finding out global patterns formed by all the training samples. At the implementation level, while the former determines the class label of a test instance by analyzing the transition probabilities or other kinds of relational information, such as neighbors' edge weights, our approach is realized by calculating the network's topological measures, permitting the extraction of some kinds of semantic structures presented in the training data.

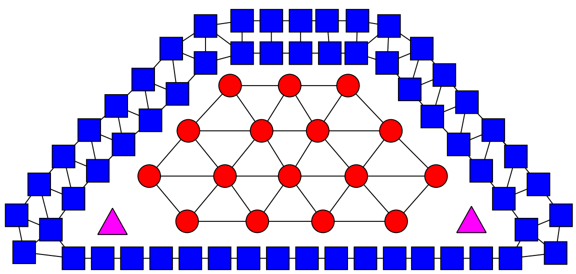

Fig. 1. A simple example of a supervised data classification task where two clear patterns are formed: the highly organized circular-shaped class and a rather dense square-shaped class. The goal is to classify the two triangle-shaped test instances.