Neurocomputing 160 (2015) 63–72

Contents lists available at ScienceDirect

Neurocomputing

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

Particle competition and cooperation for semi-supervised learning with label noise

Fabricio A. Breve [a,*], Liang Zhao [b], Marcos G. Quiles [c]

[a] *Department of Statistics, Applied Mathematics and Computation (DEMAC), Institute of Geosciences and Exact Sciences (IGCE), São Paulo State University (UNESP), Avenida 24A, 1515 – DEMAC, Bela Vista, Rio Claro, São Paulo, CEP 13506-900, Brazil*

[b] *Department of Computer Science and Mathematics (DCM), School of Philosophy, Science and Literature in Ribeirão Preto (FFCLRP), University of São Paulo (USP), Av. Bandeirantes, 3900, Monte Alegre, Ribeirão Preto, São Paulo, CEP 14040-900, Brazil*

[c] *Institute of Science and Technology (ICT), Federal University of São Paulo (Unifesp), São José dos Campos, SP, Brazil*

ARTICLE INFO

Article history:
Received 14 March 2014
Received in revised form
25 July 2014
Accepted 22 August 2014
Available online 11 February 2015

Keywords:
Label noise
Semi-supervised learning
Particle competition and cooperation

ABSTRACT

Semi-supervised learning methods are usually employed in the classification of data sets where only a small subset of the data items is labeled. In these scenarios, label noise is a crucial issue, since the noise may easily spread to a large portion or even the entire data set, leading to major degradation in classification accuracy. Therefore, the development of new techniques to reduce the nasty effects of label noise in semi-supervised learning is a vital issue. Recently, a graph-based semi-supervised learning approach based on particle competition and cooperation was developed. In this model, particles walk in the graphs constructed from the data sets. Competition takes place among particles representing different class labels, while the cooperation occurs among particles with the same label. This paper presents a new particle competition and cooperation algorithm, specifically designed to increase the robustness to the presence of label noise, improving its label noise tolerance. Different from other methods, the proposed one does not require a separate technique to deal with label noise. It performs classification of unlabeled nodes and reclassification of the nodes affected by label noise in a unique process. Computer simulations show the classification accuracy of the proposed method when applied to some artificial and real-world data sets, in which we introduce increasing amounts of label noise. The classification accuracy is compared to those achieved by previous particle competition and cooperation algorithms and other representative graph-based semi-supervised learning methods using the same scenarios. Results show the effectiveness of the proposed method.

1. Introduction

Label noise is an important issue in machine learning and, more specifically, in data classification. A classifier usually learns from a set of labeled samples to predict the classes of new samples. However, many real-world data sets contain noise, and learning from those may lead to many potential negative consequences [1]. Label noise may be of two different types: feature noise and class noise [1,2]. Feature noise affects observed values of the data features. For example, sensors may introduce some Gaussian noise during data feature measurement. On the other hand, class noise alters the labels assigned to data instances. For instance, a specialist may mistakenly assign the wrong class to some samples [3], specially when the labeling task is subjective, like in medical applications [4]. In this paper, we focus on class noise, which is potentially the more harmful type of label noise [1,2,5].

The reliability of class labels is important in supervised learning algorithms [6,7], but in semi-supervised learning this is a crucial issue. Semi-supervised learning is usually applied to problems where only a small subset of labeled samples is available, together with a large amount of unlabeled samples [8–10]. This is a common situation nowadays, as the size of the data sets being treated is constantly increasing, making prohibitive the task of labeling samples to supervised approaches. This task is time consuming and usually requires the work of human experts. Therefore, class noise is a major problem in semi-supervised learning, due to the smaller proportion of labeled data in the whole data set. In these scenarios, errors may easily affect the classification of a large portion or even the entire data set [11], leading to major degradation in classification accuracy, which is the more frequently reported consequence of label noise [1]. Therefore, it is vital to develop techniques to reduce the nasty effects of label noise in semi-supervised learning process.

[*] Corresponding author.
E-mail addresses: fabricio@rc.unesp.br (F.A. Breve), zhao@usp.br (L. Zhao), quiles@unifesp.br (M.G. Quiles).

http://dx.doi.org/10.1016/j.neucom.2014.08.082
0925-2312/© 2015 Elsevier B.V. All rights reserved.

There are three broader approaches of handling label noise in classification [12,1,13]: robust algorithms, filtering, and correction. Robust algorithms are designed to naturally tolerate a certain amount of label noise, so they do not need any special treatment. Filtering the noise means that some label noise cleaning strategy is used to identify and discard noisy labels before the training process. Finally, correction means that the noisy labels are identified, but instead of eliminating them, they are repaired or handled properly. Albeit it is not always clear whether an approach belongs to one category or the other [1]. Usually, a mixed strategy of the above-mentioned categories are used to deal with label noise problem.

Recently, a particle competition and cooperation approach was used to realize graph-based semi-supervised learning [14]. The data set is converted into a graph, where samples are nodes with edges between the similar samples. Each labeled node is associated with a labeled particle. Particles walk through the graph and cooperate with identically labeled particles to classify unlabeled samples, while competing against particles with different labels. The main advantage of particle competition and cooperation method over most other semi-supervised learning algorithms can be summarized as follows: we have proved that it has lower computational complexity [14] due to its local propagation nature; at the same time, extensive numerical studies show the method can achieve high precision of classification; it is similar to many natural or biological processes, such as resource competition by animals, territory exploration by humans (animal), election campaigns, etc. In this way, we believe that the particle competition and cooperation method can be also used back to model those natural or biological systems. The original competition and cooperation process generates much useful information and saved in the dominance level vector of each node. Such information can be used to solve other relevant problems beyond the standard machine learning tasks. For example, it can help to determine data class overlapping, fuzzy classification, and outlier detection by analyzing the distribution of the dominance vectors [15]. In this paper, we modify and further improve the original method to treat an important issue in semi-supervised learning: learning with label noise or wrong labels.

Taking the interesting features of the particle competition and cooperation approach into account, further improvements to increase the robustness of the method have been pursued. Some preliminary results were presented in [11]. The improved algorithm raised classification accuracy in the presence of label noise. However, some drawbacks have been identified, like high differences in node degree among labeled and unlabeled nodes and lack of connection between labeled particles and their corresponding labeled nodes. As a consequence, the particles spend quite more time on labeled nodes than unlabeled ones, which demands a higher number of iterations to converge. Moreover, on conditions where the amount of label noise is critical, a team of particles may switch territory with another team. This happens because particles are not strongly attracted to their corresponding nodes and they may be attracted to nodes with label noise which are on another class territory. This territory switching phenomenon always involves all particles from two or more classes, therefore it leads to major classification accuracy lost.

In this paper, we further improved the robustness of the particle competition and cooperation method to label noise. We addressed the problems of the preliminary version by enhancing graph generation, leveling nodes degrees, and thus lowering execution times. The territory switching phenomenon was also addressed by the changes in the graph generation, changes in the particles distance tables calculation, and periodic resets in particles and nodes. These improvements allow the new model to keep the particles closer to their neighborhood, increase the attraction between particles and their corresponding labeled nodes, and bring particles back after a while if they still fail to avoid territory switching eventually.

The proposed algorithm falls somewhere near the boundary between the robust algorithm approach and the correction approach aforementioned. It may be seen as a robust algorithm approach since the original algorithm has some natural tolerance to label noise, although it was not designed to handle this specific problem. In addition, it was also improved to dynamically discover and re-label label noise, thus stopping the noise propagation and allowing the algorithm to achieve higher classification accuracy. In this sense, this approach may be seem as belonging to the correction approach type. It is important to notice that this correction is a built-in feature. Both labeling unlabeled nodes and fixing label noise tasks run together in a single process.

Computer simulations presented in this paper show the effectiveness and robustness of the improved algorithm in the presence of high amounts of label noise. The classification accuracy achieved by the proposed method is compared with those achieved by all three previous versions and also with those achieved by some other representative graph-based semi-supervised learning methods [16–18]. Both artificially generated and real-world data sets were used. Label noise was introduced in these data sets with increasing levels to discover how much label noise each algorithm can handle until the classification accuracy seriously drops.

This paper is organized as follows. An overview of the particle competition and cooperation approach is shown in Section 2. The proposed model is described in Section 3. In Section 4, we present computer simulations. Finally, in Section 5 we draw some conclusions.

2. Particle competition and cooperation overview

In this section, we present an overview of the previous particle competition and cooperation models [14,19,11]. First, the vector-based data set is converted to a non-weighted and undirected graph. Each data instance becomes a graph node. Edges connecting the nodes are created according to the distance between the nodes in the data feature space. This graph generating process is described in Section 2.1. Then, a particle is created for each labeled node. Particles with the same label belong to the same team and cooperate among themselves. On the other hand, particles with different labels compete against each other. When the system runs, the particles walk in the graph, selecting the next node to visit according to the rules described in Section 2.3. Each node has a set of domination levels, one level for each class of the problem. When a particle visits a node, it will increase its class domination level on that node, at the same time that it will decrease the domination level of the other classes. Each particle possesses a strength level, which lowers or raises according to the domination level of its class in the node it is being visited. Particles also have a distance table which they update dynamically as they walk on the graph. Nodes and particles dynamics are described in Section 2.4. The stop criterion is described in Section 2.5. At the end of the iterative process, each data item is labeled after the class with the highest domination level on it.

2.1. Graph construction

Consider a vector-based data set $\chi = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^m$ with numerical attributes, and the corresponding label set $L = \{1, 2, \ldots, c\}$. The first l points $x_i (i \le l)$ are labeled as $y_i \in L$ and the remaining points $x_u (l < u \le n)$ are unlabeled, i.e, $y_u = \varnothing$. We define the graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$. $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes, where each one v_i corresponds to a sample $\mathbf{x}_i \in \chi$, and $\mathbf{E}$ is the set of edges (v_i, v_j).

In [14,19], two nodes v_i and v_j are connected if the distance (usually the Euclidean distance) between x_i and x_j is below a given threshold σ. Since the threshold may be hard to define, another option is to connect v_i and v_j if x_j is among the k-nearest neighbors