Neurocomputing 127 (2014) 43–51

Contents lists available at ScienceDirect

Neurocomputing

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

A semi-supervised classification technique based on interacting forces

Thiago H. Cupertino *, Roberto Gueleri, Liang Zhao

Institute of Mathematical Sciences and Computing, University of São Paulo, Av. Trabalhador São-carlense 400, São Carlos, São Paulo 13560-970, Brazil

ARTICLE INFO

Article history:
Received 10 January 2013
Received in revised form
2 May 2013
Accepted 31 May 2013
Available online 27 August 2013

Keywords:
Data classification
Semi-supervised learning
Label propagation
Dynamical system
Attraction forces

ABSTRACT

Semi-supervised learning is a classification paradigm in which just a few labeled instances are available for the training process. To overcome this small amount of initial label information, the information provided by the unlabeled instances is also considered. In this paper, we propose a nature-inspired semi-supervised learning technique based on attraction forces. Instances are represented as points in a k-dimensional space, and the movement of data points is modeled as a dynamical system. As the system runs, data items with the same label cooperate with each other, and data items with different labels compete among them to attract unlabeled points by applying a specific force function. In this way, all unlabeled data items can be classified when the system reaches its stable state. Stability analysis for the proposed dynamical system is performed and some heuristics are proposed for parameter setting. Simulation results show that the proposed technique achieves good classification results on artificial data sets and is comparable to well-known semi-supervised techniques using benchmark data sets.

1. Introduction

The research on new machine learning techniques and applications has been increasing more and more in diverse areas such as computer science, engineering, medical, physics, biology, etc. There are many different approaches to perform classification tasks. A traditional division is *supervised* and *unsupervised* learning paradigms [1]. Supervised learning aims at finding a rule that predicts the output of a given input data, that is, it tries to find relationships between input–output data pairs in a way that the prediction rule is more accurate as more labeled examples are given. On the other hand, the unsupervised learning paradigm seeks underlying structures in a given data set, working only with unlabeled instances.

A problem can arise when a supervised technique requires labeled instances that are hard to provide. For instance, if one wants to classify a group of web pages over the Internet according to their areas of interest, for example, news, literature, movies, sports, etc., it becomes arduous to provide many labeled examples as the Internet hosts billions of pages, and the initial categorization of each web page must be performed by a human or expert. In this case, the labeling task becomes expensive and time-consuming, a non-trivial work to perform.

A different paradigm called *semi-supervised learning* (SSL) has been studied extensively over the past years to overcome this problem. The main idea behind this paradigm is to perform classification tasks by using both: few labeled instances and the information provided by large amount of unlabeled instances [2,3]. Hence, the SSL approach could provide higher accuracies using less human efforts and exploiting the unlabeled massive group of data. This is practicable due to three SSL assumptions: manifold, smoothness and cluster [4]. The manifold assumption states that the high-dimensional data lies on a low-dimensional manifold whose properties ensure more accurate density estimation and more appropriate similarity measures. The smoothness assumption states that if two points are close to each other in a high density region, then their correspondent labels should be close to each other as well. Finally, the cluster assumption states that if two points are in the same cluster, then they are likely to be of the same class (or, in other words, to have the same label).

Many SSL algorithms have been proposed [4]. Some algorithms are mainly developed from generative models, including the Gaussian mixture model [5], mixture of experts [6] and extensions [7,8], transductive and semi-supervised support vector machines [9,10], and boosting algorithms [11,12]. Also, co-training is another important methodology [13]. Some techniques are graph-based [3,14]. These techniques basically map the data instances in an underlying graph and then uses the graph structure to perform the classification task. In some approaches, graph nodes represent states, and links between nodes represent transition probabilities in a random walk process [15,16]. As mentioned in [4], an important problem in this paradigm is the model correctness, that is, unlabeled data may decrease the accuracy with an incorrect model assumption.

In this paper, we propose a nature-inspired SSL technique based on attraction forces. It models data instances as points in a k-dimensional space and performs their motion according to the resultant force applied upon them. The labeled instances act as attraction points while the unlabeled instances receive the forces

* Corresponding author. Tel.: +55 16 3373 8161.
E-mail addresses: thcupertino@gmail.com, thiagohc@icmc.usp.br (T.H. Cupertino).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.050

and move towards the attraction points. At a certain moment of the dynamics, the unlabeled instances receive a label that is propagated from the labeled points and become new attraction points. In spite of its simplicity, the model is effective and provides good classification results.

The theoretical basis of the technique presented in this paper was originally published in [17]. Here, we provide an extension of that research including some relevant contributions: (1) We have improved the proposed model. Specifically, we propose the usage of a precision value to scale the movements of instances during the system dynamics to normalize the system and to achieve a more robust convergence. (2) We have expanded the numerical analysis largely. In this case, extensive studies on the influence of parameter values and their combinations to illustrate optimal regions in both artificial and real data sets are provided.

This paper is organized as follows: Section 2 introduces the technique and its mathematical modeling. Section 3 analyzes the stability of the proposed dynamical system. Section 4 presents heuristics used for parameter selection. Finally, Section 5 provides simulation results and discussions, and Section 6 concludes the paper.

2. Proposed technique

The use of attraction forces between labeled and unlabeled instances can provide a model for SSL that fits well into the smoothness and cluster assumptions. Labeled instances are considered as fixed attraction points that apply attraction forces on the unlabeled instances. In turn, the later are expected to move towards the resultant force direction and, eventually, to converge to an attraction point. Once close enough, say inside a neighborhood region δ, the label from the attraction point propagates to the unlabeled neighbor, and it becomes a new fixed attraction point. At the end of the process it is expected that all points converge to some attraction point. By means of the attraction forces, instances are kept together in their dense groups (clusters), while different labeled points are responsible for dividing the space under the smoothness assumption.

This process uses the initial labeled instances information to propagate their labels to the attracted instances which, in turn, after being labeled, propagates it to their attracted instances and so on. In other words, this dynamic makes use of unlabeled data (the attracted neighbors) information to perform the classification task which is, in turn, the main idea of a SSL technique. Even more, the model allows fine adjustment as one can use any attraction force function and adjust its parameters.

Despite the simplicity of the model, two considerations are necessary to accomplish the above mentioned behavior and classify the unlabeled instances correctly. One of them is to guarantee that the process is stable, and the other is to certify that the labels propagate adequately through the unlabeled instances, in the sense that the algorithm will converge and achieve good classification accuracy. The stability issue can be treated using similar approaches from swarm aggregation works [18,19], while the label propagation dynamics can be analyzed in terms of the attraction force function parameters. Both are explained in the next sections.

2.1. Mathematical modeling

We consider that it is given a data set $\mathcal{D}=\{\mathcal{L}\cup\mathcal{U}\}$ composed of sets of labeled $\mathcal{L}=\{\mathbf{x}_i^{(l)}, i=1,\ldots,n\}$ and unlabeled $\mathcal{U}=\{\mathbf{x}_i, i=1,\ldots,m\}$ instances. The set of labels form a finite set $\mathcal{B}$ ($l\in\mathcal{B}$), and each instance is represented by k attributes: $\mathbf{x}_i=\{x_{i1},x_{i2},\ldots,x_{ik}\}$. The objective is to classify all instances in $\mathcal{U}$ in a transductive way, by using the labeled instances in $\mathcal{L}$. The instances are modeled as points, ignoring their dimensions. We assume synchronous motion and no time delays, that is, all points move simultaneously and know the exact position of each other. The motion of unlabeled points $\mathbf{x}_i$ is governed by the following system:

$$\dot{\mathbf{x}}_i(t)=\sum_{j=1,j\neq i}^{n} f[\mathbf{x}_j^{(l)}(t)-\mathbf{x}_i(t)],\quad i=1,\ldots,m, \tag{1}$$

where function f is the attraction force among instances. As described by Eq. (1), each unlabeled instance $\mathbf{x}_i$ receives attractive forces from all labeled instances and the resultant force is the sum of all individual forces. Thus, the direction and magnitude of $\mathbf{x}_i(t)$'s motion is determined by the forces applied by the labeled instances.

The attraction function is defined as a Gaussian field with parameters α and β:

$$f[\mathbf{x}_j^{(l)}(t)-\mathbf{x}_i(t)]=[\mathbf{x}_j^{(l)}(t)-\mathbf{x}_i(t)]\frac{\alpha}{e^{\beta\|(\mathbf{x}_j^{(l)}(t)-\mathbf{x}_i(t))\|^2}}. \tag{2}$$

We choose an attraction function in order to guarantee that the more a point is close to an attractor point, the more strong is the force. Moreover, its parameters provide an easy way to adjust the function amplitude and range, which is necessary to the correct functioning of the process. In the next subsections, the heuristics used to adjust these parameters are described in detail.

2.2. Summarized algorithm

In a concise form, the proposed technique can be summarized by Algorithm 1. The technique is performed iteratively in four steps (from 2 to 5), until all instances are labeled. The parameter initialization (step 1) is discussed in Section 4.

Algorithm 1. Proposed technique.

Input:
$\mathcal{L}$: labeled data set
$\mathcal{U}$: unlabeled data set
Output:
l_i: estimated class for each $\mathbf{x}_i\in\mathcal{U}$
Initialization:
1. $(\alpha,\beta,\delta)=$ Initialize parameters
Classification:
DO
2. Calculate distances among points
3. Calculate attraction forces
4. Update points' positions
5. Update labels
WHILE (there are unlabeled instances)

3. Stability analysis

For the sake of completeness, the stability of the system in Eq. (1) is analyzed using the Lyapunov stability method [20]. Based on this method, a system $f(x(t))$ is called stable if there is a candidate function $V(x)\geq 0$ positive definite, that is, $V(x)=0$ if and only if $x=0$, and its derivative $\dot{V}(x)=(d/dt)V(x)\leq 0$ is negative definite, that is, the equality holds if and only if $x=0$. The problem is to find a suitable candidate function so that the above constraints are satisfied.

Given that in the proposed system the labeled instances are fixed, they do not receive any attraction force and so do not move, we turn our attention to the unlabeled points, which compose the system dynamics. Consider an unlabeled point $\mathbf{x}_i$ has been attracted by the resultant force function in the direction of a specific labeled point $\mathbf{x}_p^{(l)}$. In this case, $\mathbf{x}_i$ will putatively enter into $\mathbf{x}_p^{(l)}$ neighborhood δ and become labeled. Using the difference variable $e_i(t)=\mathbf{x}_i(t)-\mathbf{x}_p^{(l)}$, the