Neurocomputing 177 (2016) 385–393

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised learning via mean field methods

Jianqiang Li [a,b,d,*], Fei Wang [c]

[a] School of Software Engineering, Beijing University of Technology, China.
[b] Beijing Engineering Research Center for IoT Software and Systems, China
[c] Department of Computer Science and Engineering, University of Connecticut, United States
[d] Guangdong Key Laboratory of Popular High Performance Computers, Shenzhen Key Laboratory of Service Computing and Applications, China

ARTICLE INFO

Article history:
Received 8 February 2015
Received in revised form
16 May 2015
Accepted 15 November 2015
Communicated by Feiping Nie
Available online 26 November 2015

Keywords:
Mean field
Semi-supervised learning

ABSTRACT

The recent years have witnessed a surge of interest in *semi-supervised learning* methods. Numerous methods have been proposed for learning from partially labeled data. In this paper, a novel semi-supervised learning approach based on statistical physics is proposed. We treat each data point as an *Ising* spin and the interaction between pairwise spins is captured by the similarity between the pairwise points. The labels of the data points are treated as the directions of the corresponding spins. In semi-supervised setting, some of the spins have fixed directions (which corresponds to the labeled data), and our task is to determine the directions of other spins. An approach based on the *Mean Field* theory is proposed to achieve this goal. Finally the experimental results on both toy and real world data sets are provided to show the effectiveness of our method.

1. Introduction

Traditional data mining approaches can be categorized into two categories: (1) *supervised learning*, which aims to predict the labels of any new data points from the observed data-label pairs. A supervised learning task is called *regression* when the predicted labels take real values, and *classification* when the predicted labels take a set of discrete values. Typical supervised learning methods include *Support Vector Machine* [32] and *Decision Tree* [25]; (2) *unsupervised learning*, the goal of which is just to organize the observed data points without labels. Typical unsupervised learning tasks include *clustering* [15] and *dimensionality reduction* [26].

In this paper, we will focus on *classification*, which is traditionally a *supervised learning* task. To train a classifier one needs a collection of labeled data points. However, in many practical applications of pattern classification and data mining, one often faces a lack of sufficient labeled data, since labeling often requires expensive human labor and much time. Meanwhile, in many cases, large numbers of unlabeled data can be far easier to obtain. For example, in text classification, one may have an easy access to a large database of documents (e.g. by crawling the web), but only a small part of them are classified by hand.

Consequently, *semi-supervised learning methods*, which aim to learn from partially labeled data, are proposed. Many conventional semi-supervised learning algorithms adopt a generative model for the classifier and employ *Expectation Maximization* (*EM*) [10] to model the label prediction or parameter estimation process. For example, *mixture of Gaussians* [27], *mixture of experts* [20], and *naive Bayes* [21] have been respectively used as the generative model, while EM is used to combine labeled and unlabeled data for classification. There are also many other algorithms such as *co-training* [6], *transductive SVM* (*TSVM*) [14], and the *Gaussian process* approach [18]. For a detailed literature survey one can refer to [36].

The basic assumption behind semi-supervised learning is the *cluster assumption* [8], which states that two points are likely to have the same class label if there is a path connecting them passing through the regions of high density only. Zhou et al. [34] further explored the geometric intuition behind this assumption: (1) nearby points are likely to have the same label; (2) points on the same structure (such as a cluster or a submanifold) are likely to have the same label. Note that the first assumption is local, while the second one is global. The cluster assumption implies us to consider both local and global information contained in the dataset during learning.

In recent years there has been significant interest in adapting numerical [22] and analytic [3] techniques from statistical physics to provide beautiful algorithms and estimates for machine learning and neural computation problems. In this paper we formulate the problem of *semi-supervised learning* as that of measuring

* Corresponding author at: School of Software Engineering, Beijing University of Technology, China.
E-mail address: lijianqiang@bjut.edu.cn (J. Li).

http://dx.doi.org/10.1016/j.neucom.2015.11.042
0925-2312/

equilibrium properties of an *homogeneous Ising* model. In our model, each data point is viewed as a spin, the direction of the spin stands for the label of the data point. We also introduce some *interactions* between pairwise points based on the intrinsic geometry of the dataset. The directions of the spins corresponding to the labeled data points are fixed. And our goal is to predict the labels of the unlabeled points, which will be estimated by the directions of these spins in thermal equilibrium. The experiments show that our method can give good classification results.

The rest of this paper is organized as follows. The detailed description of the Ising model will be presented in Section 2. In Section 3 we will introduce a *Mean Field* approach for solving the Ising problems. Our approach for semi-supervised learning will be described in Section 4, and we also compare it with traditional *Bayesian* methods in Section 5. The experimental results on both toy and real world datasets will be introduced in Section 6, followed by the conclusions and discussions in Section 7.

2. Ising model

The *Ising model* [12] first proposed by E. Ising is a lattice model, which is used for describing intermolecular forces. For example, in magnets, each molecule has a *spin* that can be oriented either *up* or *down* relative to the direction of an externally applied field.

Fig. 1 shows us such an example which is a 2D periodic lattice having an array of 25 fixed points. Note that in real world applications the lattice can be of any type. With each lattice site is associated a spin variable S_i being either $+1$ or -1. In magnets, we usually call $+1$ *spin up* and -1 *spin down*.

A *configuration* of the lattice is a particular set of values of all spins, e.g. a configuration of the 5×5 regular lattice is illustrated in Fig. 1. Clearly, for such a spin system there are totally 2^{25} possible configurations.

We usually assign an energy to a specific configuration of an Ising model, and assume that the molecules exert only short-range forces on each other, i.e. the interaction energy depends only on the configurations of neighboring spins of the lattice. Taking *spontaneous magnetization* as an example, since the neighboring spins tend to have the same direction, we can define that if the two neighboring spins are in the same direction, the energy between them is $-U$, if they are in different directions, the energy is $+U$ ($U>0$ is a constant). Then the energy of the system shown in Fig. 1 is $-24U$.

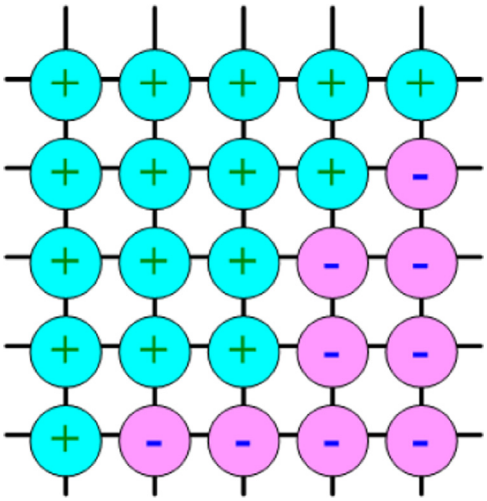

Fig. 1. An example of the 2D Ising model.

In addition, we also assume that the total energy of a configuration also includes a term θ_i for each spin, which can be viewed as the effects of an external field. If it is a magnetic field that can result in a $+\theta$ energy on the spins with $+1$ states, and $-\theta$ energy on -1 spins, then the total energy of the configuration shown in Fig. 1 is $-24U+3\theta$.

Based on the discussions above, we can define the energy of a general Ising model in a given configuration $\mathbf{S}=(S_1,S_2,\ldots,S_N)$ to be

$$\tilde{E}(\mathbf{S})=-\sum_{\langle i,j\rangle}\tilde{J}_{ij}S_iS_j-\sum_i\tilde{\theta}_iS_i, \tag{1}$$

where $S_i\in\{+1,-1\}$ is the current value of the ith spin, $\langle i,j\rangle$ represents a neighboring spin pair, J_{ij} is the symmetric *interaction energy* of the pairwise spins i and j, $\tilde{\theta}_i$ is the energy on spin i brought by the *external fields*. The *canonical partition function* of the system is defined as[1]

$$Z=\int dS_1\int dS_2\cdots\int dS_N e^{-\beta\tilde{E}(\mathbf{S})}, \tag{2}$$

where

$$\beta=(kT)^{-1}, \tag{3}$$

and k is the *Boltzmann constant* and T is the temperature. We further define the energy function as

$$E(\mathbf{S})=-\sum_{\langle i,j\rangle}J_{ij}S_iS_j-\sum_i\theta_iS_i, \tag{4}$$

where

$$J_{ij}=\beta\tilde{J}_{ij},\quad \theta_i=\beta\tilde{\theta}_i. \tag{5}$$

Then the probability distribution of the spin system is

$$P(\mathbf{S})=\frac{1}{Z}\exp\left[\left(\sum_{\langle i,j\rangle}J_{ij}S_iS_j+\sum_i\theta_iS_i\right)\right]. \tag{6}$$

Furthermore, by absorbing the constraint that S_i can only take binary values, we can rewrite $P(\mathbf{S})$ as

$$P(\mathbf{S})=\frac{\rho(\mathbf{S})}{Z}\exp\left[\left(\sum_{\langle i,j\rangle}J_{ij}S_iS_j+\sum_i\theta_iS_i\right)\right], \tag{7}$$

where $\rho(\mathbf{S})=\prod_i\rho(S_i)$ with

$$\rho(S_i)=\frac{1}{2}\delta(S_i-1)+\frac{1}{2}\delta(S_i+1), \tag{8}$$

which states the prior knowledge that each spin has the equal probabilities to be $+1$ or -1, and $\delta(\cdot)$ is the *Dirac Delta function*. Therefore the marginal probability of S_i is

$$P(S_i)=\int\prod_{j\neq i}dS_j\,P(\mathbf{S}). \tag{9}$$

Our goal is to approximate the behavior of such an Ising type interacting spin system in equilibrium. We will introduce an *adaptive TAP* approach [22] in the next section to solve the problem.

3. The mean field approach for Ising model

The main idea of the *mean field* theory is to focus on one spin and assume that the most important contribution to the interactions of such spin with its neighboring spins is determined by the *mean field* due to its neighboring spins [24]. It originally aims to

[1] Here we give a more general form of the partition function. In our Ising model case, since each random variable can only have two integer values, we can use the sum operator to replace the integral operator.