Semi-Supervised Mean Fields

Fei Wang **Shijun Wang** **Changshui Zhang** **Ole Winther**
State Key Lab of Intelligent Technologies and Systems Intelligent Signal Processing
Department of Automation Informatics and Mathematical Modelling
Tsinghua University, Beijing, P.R.China Technical University of Denmark

Abstract

A novel semi-supervised learning approach based on statistical physics is proposed in this paper. 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

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 [5][25]. The basic assumption behind semi-supervised learning is the *cluster assumption* [6], 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.* [23] 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 [15] and analytic [1] 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 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* [10] first proposed by E. Ising is a lattice model, which is used for describing intermolecular forces. The lattice can be of any type. For example, in magnets, each molecule has a *spin* that can be ori-

ented either *up* or *down* relative to the direction of an externally applied field.

A *configuration* of the lattice is a particular set of values of all spins, *e.g.* the number of different configurations of the 5x5 regular lattice is 2^{25}.

We usually assign an energy to a specific configuration of an Ising model, typically the energy of a general Ising model in a given configuration $\mathbf{S} = (S_1, S_2, \cdots, S_N)$ to be

$$\tilde{E}(\mathbf{S}) = -\sum_{<i,j>} \tilde{J}_{ij} S_i S_j - \sum_i \tilde{\theta}_i S_i, \tag{1}$$

where $S_i \in \{+1, -1\}$ is the current value of the i-th spin, $< i, j >$ 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

$$E(\mathbf{S}) = -\sum_{<i,j>} J_{ij} S_i S_j - \sum_i \theta_i S_i, \tag{4}$$

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

$$P(\mathbf{S}) = \frac{1}{Z} e^{\sum_{<i,j>} J_{ij} S_i S_j + \sum_i \theta_i S_i}, \tag{6}$$

where S_i $(i = 1, 2, \cdots N)$ are the random variables which can only take value +1 or -1. Therefore the marginal probability of S_i is

$$P(S_i) = \int \prod_{j \neq i} dS_j P(\mathbf{S}). \tag{7}$$

Our goal is to approximate the behavior of such an Ising type interacting spin system in equilibrium. We will introduce an *Naïve Mean Field* approach [15] in the next section to solve the problem.

[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.

3 The Naïve 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 [16]. It originally aims to approximate the behavior of interacting spin systems in thermal equilibrium. We use it here to approximate the behavior of Ising models in equilibrium. More concretely, the mean state of spin S_i is

$$\langle S_i \rangle = \int dS_i P(S_i), \tag{8}$$

where $P(S_i)$ is the marginal distribution in Eq.(7). So we can define the *mean field* $\langle \mathbf{S} \rangle = (\langle S_1 \rangle, \langle S_2 \rangle, \cdots, \langle S_N \rangle)$ by these mean values. However, in most of the cases, the exact form of $P(\mathbf{S})$ is not available. So the computation of the exact means of S_i $(i = 1, 2, \cdots, N)$ become intractable. To solve such a problem, we can first approximates $P(\mathbf{S})$ by $Q(\mathbf{S})$ which belongs to a family $\mathcal{M}$ of easily tractable distributions [15]. The $Q(\mathbf{S})$ is chosen to minimize the following *Kullback-Leibler divergence*

$$KL(Q\|P) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{Q}{P}. \tag{9}$$

Bringing Eq.(6) into the above equation we can get

$$KL(Q\|P) = \ln Z + V(Q) - S(Q), \tag{10}$$

where

$$S(Q) = -\sum_{\mathbf{S}} Q(\mathbf{S}) \ln Q(\mathbf{S}) \tag{11}$$

is the *entropy* of the distribution Q, and

$$V(Q) = \sum_{\mathbf{S}} Q(\mathbf{S}) E(\mathbf{S}) \tag{12}$$

is the *variational energy*. The *mean field approximation* can be obtained by first approximate the distribution family $\mathcal{M}$ by all product distributions

$$Q(\mathbf{S}) = \prod_j Q_j(S_j). \tag{13}$$

For *Ising models*, the most general form of the Q_j's is

$$Q_j(S_j; m_j) = \frac{1 + S_j m_j}{2}, \tag{14}$$

where $m_j = \langle S_j \rangle_Q$, and $\langle \cdot \rangle_Q$ represents the mean over distribution Q. Then the variational entropy of Q is

$$S(Q) = -\sum_i \left(\frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right),$$

Table 1: The NMF Method For Ising Model.

Initialization:
- Start from a *tabular rasa* $\langle \mathbf{S} \rangle = (\langle S_1 \rangle, \langle S_2 \rangle, \cdots, \langle S_N \rangle) = \mathbf{0}$ (or small values if $\langle \mathbf{S} \rangle = \mathbf{0}$ is a fixed point).
- Learning rate $\eta = 0.05$.
- Fault tolerance $ft = 10^{-3}$.

Iterate:

do:

for all i**:**

Compute m_i by Eq.(18).

$\delta \langle S_i \rangle = m_i - \langle S_i \rangle$

end for

for all i**:**

$\langle S_i \rangle = \langle S_i \rangle + \eta \delta \langle S_i \rangle$

end for

while $\max_i |\delta \langle S_i \rangle|^2 > ft$

and the variational energy reduces to

$$V(Q) = \langle E(\mathbf{S}) \rangle_Q = -\sum_{<i,j>} J_{ij} m_i m_j - \sum_i m_i \theta_i. \quad (15)$$

Hence, according to Eq.(10), what we need to do is only to minimize[2]

$$F(Q) = V(Q) - S(Q). \quad (16)$$

Then

$$\frac{\partial F(Q)}{\partial m_i} = -\sum_j J_{ij} m_j - \theta_i + \frac{1}{2} \ln \frac{1+m_i}{1-m_i}. \quad (17)$$

Setting $\frac{\partial F(Q)}{\partial m_i} = 0$ we can easily get the *mean field equations*

$$m_i = \tanh \left(\sum_j J_{ij} m_j + \theta_i \right). \quad (18)$$

In such a way, the intractable task of computing the exact averages over P is replaced by the solution of Eq.(18). An iterative method for achieving this goal is shown in Table 1.

4 Semi-Supervised Learning Based on the Mean Field Approach

So far we have defined the Ising model and introduced a mean field approach for solving the states of the spins in such a model in thermal equilibrium. We can now turn to the problem for which these concepts will be utilized: *semi-supervised learning*.

In semi-supervised learning, the dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_n\}$ is composed of two parts. The first part $\mathcal{X}_L = \{\mathbf{x}_i\}_{i=1}^l$ is the labeled set, in which the data points are labeled with $+1$ or -1 (we consider the two-class case for the moment). The remaining unlabeled data points $\mathcal{X}_U = \{\mathbf{x}_u\}_{u=l+1}^n$ constitute the second part of $\mathcal{X}$. The goal of semi-supervised learning is to predict the labels of the unlabeled data[3].

Considering the two-class classification problem, we can just treat each data point as a spin, and its label as the direction of the spin. Then the labels of the unlabeled data can be regarded as the state of those spins in equilibrium, *i.e.* the spins $\{S_i\}_{i=1}^N$ can be viewed as the *hidden label variables* of the data. So there are two things remained: (1) how to determine the interactions between pairwise data (*i.e.* $\tilde{J}_{ij}$); (2) how to determine the external field (*i.e.* $\tilde{\theta}_i$).

4.1 Compute the Spin-Spin Interactions

No doubtedly, the computation of the spin-spin interactions is at the heart of our semi-supervised algorithm. An intuition here is that the more similar $\mathbf{x}_i$ to $\mathbf{x}_j$, the stronger the interaction between $\mathbf{x}_i$ and $\mathbf{x}_j$ will be. Therefore, we should define a proper similarity between pairwise data. Some possible ways for defining such a similarity including:

1. *Unweighted k-Nearest Neighborhood Similarity* [2]: The similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ is 1 if $\mathbf{x}_i$ is in the k-nearest neighborhood of $\mathbf{x}_j$ or $\mathbf{x}_j$ is in the k-nearest neighborhood of $\mathbf{x}_i$, and 0 otherwise. k is the only hyperparameter that controls this similarity. As noted by [27], this similarity has the nice property of "adaptive scales", since the similarities between pairwise points are the same in low and high density regions.

2. *Weighted Linear Neighborhood Similarity* [22]: Assuming $\mathbf{x}_i$ is in the neighborhood (k-nearest neighborhood or ϵ-ball neighborhood) of $\mathbf{x}_j$, then the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, w_{ij}, can be computed by solving the following optimization prob-

[2]Since Z is not dependent on Q

[3]The task we want to handle here is also called transduction, in which we do not consider to predict the labels of the data not in the training dataset (induction). Since according to [21], this may import unnecessary complexity to the learning algorithm.

lem

$$\begin{aligned} min_{w_{ij}} \quad & \|\mathbf{x}_i - w_{ij}\mathbf{x}_j\|^2 \\ s.t. \quad & w_{ij} \geqslant 0, \quad \sum\nolimits_j w_{ij} = 1, \end{aligned}$$

and $w_{ij} = 0$ if $\mathbf{x}_j$ is not in the neighborhood of $\mathbf{x}_i$. Note that this similarity is generally asymmetric, and we can use $\tilde{w}_{ij} = \frac{1}{2}(w_{ij} + w_{ji})$ for symmetrize it. The neighborhood size (k or ϵ) is the only hyperparameter for controlling this similarity.

3. *Weighted Exponential Similarity* [2][7][23][27]: Let d_{ij} be the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, then the tanh similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be computed by

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma}\right), \tag{19}$$

which is also a continuous weighting scheme with σ controlling the decay rate.

Because of its popularity and solid theoretical foundations [3], we adopt the *Weighted Exponential Similarity* as our similarity measure for computing the pairwise similarities (interactions). Then another problem arises, *i.e.* defining a proper distance function. We also list some possible distance functions here.

1. *Euclidean Distance* [2][23][27]: The distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be calculated by

$$d_{ij}^E = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}. \tag{20}$$

2. *Connectivity Distance* [9]: We also need to first construct a *connected neighborhood graph* for the dataset $\mathcal{X}$. Let p be a path from one point to another with length $|p|$, and the indices of the data points on this path are denoted by $\{p_k\}_{k=1}^{|p|}$. Let $\mathcal{P}_{ij}$ be the set of paths connecting $\mathbf{x}_i$ and $\mathbf{x}_j$, then the *connectivity distance* between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined by

$$d_{ij} = \min_{p \in \mathcal{P}_{ij}} \max_{1 \leqslant k \leqslant |p|-1} d^E_{p_k p_{k+1}},$$

where d^E represents the Euclidean distance. Chapelle *et al* [7] further propose to "soften" this distance to make it more robust to the *bridge points*. Such "softened" connectivity distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed by

$$d_{ij} = \frac{1}{\rho^2} \ln\left(1 + \min_{p \in \mathcal{P}_{ij}} \sum_{k=1}^{|p|-1} \left(e^{\rho d^E_{p_k p_{k+1}}} - 1\right)\right),$$

where $\rho > 0$ is a free parameter for controlling the distance.

3. *Inner Product Distance* [23][25]: The inner product distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed by

$$d_{ij} = 1 - \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}. \tag{21}$$

Note that this distance is one of the most commonly used distances in text classification.

In real world applications, using which distance is dependent on our prior knowledge of the datasets. For example, for the data (nearly) residing on intrinsic manifolds, then the *geodesic distance* can be a better choice; for the dataset that is corrupted by some noise, then using *connectivity distance* might be better; for text analysis, the *inner product* distance should be considered for the first choice. However, if we do not have any prior information about the dataset, then the *Euclidean distance* may be a safe choice.

4.2 Determining the External Fields

Another important issue that we should address in our method is how to determine the influence of the external fields, *i.e.* $\tilde{\theta}_i$ ($i = 1, 2, \cdots, N$). Recall that what we want to solve is a semi-supervised learning problem, that means some data have already been labeled. Considering an analogy to the spin system, we can regard the labeled points as the spins having "fixed" directions, which is forced by some external fields.

For a particular configuration of the spin system, if the states for the spins corresponding to the labeled data points are in accordance with their imposed states (labels), then its probability should be higher. Recalling the definition of the probability of a specific configuration in Eq.(6), as for each spin there are only two possible directions, *spin up* (+1), or *spin down* (-1), we can define $\tilde{\theta}_i$ in the following way

$$\tilde{\theta}_i = \begin{cases} l_i, & if \ \ \mathbf{x}_i \ \ is \ \ labeled \ \ as \ \ l_i \\ 0, & if \ \ \mathbf{x}_i \ \ is \ \ unlabeled \end{cases}. \tag{22}$$

In such a definition, if the state of a "labeled" spin is the same as its label, then $\tilde{\theta}_i > 0$, which will result in a higher probability, otherwise $\tilde{\theta}_i < 0$, which will make the probability lower. And for simplicity, we do not consider the influence of the external fields on unlabeled spins.

4.3 Algorithm Framework

Having computed the *spin-spin interactions* and the *influence of the external fields*, we can derive the framework of our algorithm, which is shown in Table 2. There are some issues should be addressed here:

Table 2: Semi-Supervised Learning via NMF.

Inputs:
- Dataset $\mathcal{X}$, Scale σ, Temperature T
- A proper distance function $d(\cdot,\cdot)$

Outputs:
- The labels of the unlabeled data.

1. Calculate the distance matrix $\mathbf{D}$ with its (i,j)-th entry $D_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$;
2. Calculate the interaction matrix $\tilde{\mathbf{J}}$ with its (i,j)-th entry $\tilde{J}_{ij} = \exp(-D_{ij}/\sigma)$;
3. Normalize $\tilde{\mathbf{J}}$ by $\tilde{\mathbf{J}} = \mathbf{R}^{-1/2}\tilde{\mathbf{J}}\mathbf{R}^{-1/2}$, where $\mathbf{R}$ is a diagonal matrix with the i-th element on its diagonal line $R_{ii} = \sum_j \tilde{J}_{ij}$;
4. Compute the matrix $\mathbf{J}$ with its (i,j)-th entry $J_{ij} = \beta\tilde{J}_{ij}$, where β is defined as in Eq.(3),
5. Calculate $\langle S_u \rangle$ $(\mathbf{x}_u \in \mathcal{X}_U)$ by the method shown in Table 1, and fix $\langle S_l \rangle$ $(\mathbf{x}_l \in \mathcal{X}_l)$ to be t_l, which is the label of $\mathbf{x}_l$;
6. If $\langle S_i \rangle > 0$, then classify $\mathbf{x}_i$ as +1, else if $\langle S_i \rangle < 0$, classify $\mathbf{x}_i$ as -1

1. The normalization procedure in step 3 is to make the computed similarities insensitive to the data distribution, since a single Gaussian function will always assign high similarities to high density regions, which will bias the final classification results if the data from different classes have different densities [24].

2. Usually the computed $\langle S_i \rangle$ is not just +1 or −1, so we classify the data points as the sign of $\langle S_i \rangle$. And $\langle S_i \rangle$ can be viewed as the *soft label* of $\mathbf{x}_i$.

3. For multi-class problems, we can just use (1) *one-vs-rest* scheme [21] and classify $\mathbf{x}_i$ to the class with the largest S_i value; (2) a *Potts* multi-valued spin model.

5 Relationship with Bayesian Discriminative Methods

The *Bayesian approach* is an importance class of methods for data mining and machine learning. Traditionally, a *Bayesian classifier* can be categorized into either *generative* or *discriminative* classifiers. The goal of *generative classifiers* is to learn a joint probability model, $P(\mathbf{x}, t)$, of input $\mathbf{x}$ and its class label t, and then making their predictions by using the *Bayesian rule* to compute $P(t|\mathbf{x})$. *Discriminative classifiers*, on the other hand, model posterior class probabilities $P(t|\mathbf{x})$ for all classes directly and learn a mapping from $\mathbf{x}$ to t. It has often been argued that for many application domains, *discriminative classifiers* can often achieve higher test accuracies than *generative classifiers*[21].

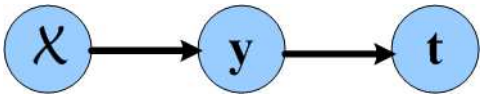


Figure 1: The graphical model of a discriminative classifier.

In the following we will use the same symbol system introduced at the beginning of section 4. We use $\mathbf{t} = (t_1, t_2, \cdots, t_n)$ to denote the *hard labels* of the data points, *i.e.* $t_i \in \{-1, +1\}$, $\forall\ 1 \leqslant i \leqslant n$, and $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ is the *hidden soft labels* of the data points, *e.g.* in the two-class classification problem, $\mathbf{t} = sign(\mathbf{y})$. In the *discriminative framework*, we can draw a *graphical model* for such a problem which is shown in Fig.1.

For semi-supervised classification, what we observed is the set $\mathcal{D} = \{\mathcal{X} = \mathcal{X}_L \bigcup \mathcal{X}_U, \mathbf{t}_L\}$, where $\mathbf{t}_L$ corresponds to the labels of the labels of the labeled points. And we aim at learning the posterior $P(\mathbf{t}_U|\mathcal{D})$, where $\mathbf{t}_U = (t_{l+1}, t_{l+2}, \cdots, t_n)$. The posterior can be written as

$$P(\mathbf{t}_U|\mathcal{D}) = \int_{\mathbf{y}} P(\mathbf{t}_U|\mathbf{y})P(\mathbf{y}|\mathcal{D}) \tag{23}$$

The same as in [12], we can first approximate $P(\mathbf{y}|\mathcal{D})$ and then use Eq.(23) to classify the unlabeled data. Using the *Bayesian rule*, $P(\mathbf{y}|\mathcal{D})$ can be defined as

$$P(\mathbf{y}|\mathcal{D}) = P(\mathbf{y}|\mathcal{X}, \mathbf{t}_L) = \frac{1}{Z}P(\mathbf{y}|\mathcal{X})P(\mathbf{t}_L|\mathbf{y})$$

The term, $P(\mathbf{y}|\mathcal{X})$ is the probability of the hidden labels given the training set, which can be regarded as the prior information of $\mathbf{y}$ on the training set. $P(\mathbf{t}_L|\mathbf{y})$ is the *likelihood term*, which is usually written as $P(\mathbf{t}_L|\mathbf{y}) = \prod_{i=1}^{l} P(t_i|y_i)$ [12]. This term models the probabilistic relation between the observable hard labels and the hidden soft labels.

The *prior term* plays an important role in semi-supervised learning, especially when the size of the labeled data is small. The recent research shows that the prior should impose a smoothness constraint with respect to the intrinsic data structure [12][18], which means that it should give higher probability to the labelings that respect the similarity of the data graph.

More concretely, if we model the whole dataset as a *weighted undirected* graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{X}$ is the node set, and $\mathcal{E}$ is the edge set. Associated with each edge $e_{ij} \in \mathcal{E}$ is a symmetrical nonnegative weight representing the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. In this way, the intrinsic structure of the dataset can be described by a *data graph*, and the smooth labeling with

respect to the data structure just corresponds to the smooth labeling over the data graph. As noted in [3], if we regard $\mathbf{y}$ as a function defined on the data graph $\mathcal{G}$, such that y_i is the return value of this function at $\mathbf{x}_i$, then the smoothness of $\mathbf{y}$ can be computed by

$$\mathcal{S}_{\mathbf{y}} = \mathbf{y}^T \mathbf{M} \mathbf{y}, \tag{24}$$

where $\mathbf{M}$ is the so-called *smoothness matrix*, some typical choice of $\mathbf{M}$ are

- *Combinatorial Graph Laplacian* [3][27]

$$\mathbf{M} = \mathbf{R} - \tilde{\mathbf{J}} \tag{25}$$

- *Normalized Graph Laplacian* [23][24]

$$\mathbf{M} = \mathbf{I} - \mathbf{R}^{-1/2} \tilde{\mathbf{J}} \mathbf{R}^{-1/2} \tag{26}$$

where $\tilde{\mathbf{J}}$ is the similarity matrix and $\mathbf{R}$ is the diagonal matrix with its (i,i)-th element equal to the sum of the i-th row of $\tilde{\mathbf{J}}$.

It is found that using the *normalized graph Laplacian* can usually produce better results [23]. Therefore, a natural choice for the prior of $\mathbf{y}$ could be

$$P(\mathbf{y}|\mathcal{X}) = \frac{1}{Z_p} \exp\left(-\mathbf{y}^T (\mathbf{I} - \mathbf{R}^{-1/2} \tilde{\mathbf{J}} \mathbf{R}^{-1/2}) \mathbf{y}\right).$$

The likelihood term $P(\mathbf{t}_L|\mathbf{y})$ should be defined in the way that it should be higher when t_i and y_i have the same sign, and lower otherwise. Therefore a natural choice for the likelihood term could be

$$P(\mathbf{t}_L|\mathbf{y}) = \prod_{i=1}^{l} P(t_i|y_i) = \frac{\exp\left(\sum_{i=1}^{l} t_i y_i\right)}{\prod_{i=1}^{l} \left(\exp(y_i) + \exp(-y_i)\right)}$$

Note that for the labeled points, their *hard labels* are just the influence of the external fields, *i.e.* $\{\tilde{\theta}_i\}_{i=1}^{l}$, as we defined in Eq.(22). As a result, the posterior probability

$$P(\mathbf{y}|\mathcal{D}) = \frac{\frac{1}{Z} e^{-\mathbf{y}^T \mathbf{y} + \mathbf{y}^T \mathbf{R}^{-1/2} \tilde{\mathbf{J}} \mathbf{R}^{-1/2} \mathbf{y} + \sum_{i=1}^{l} \tilde{\theta}_i y_i}}{\prod_{i=1}^{l} \left(e^{y_i} + e^{-y_i}\right)}.$$

Since we can always normalize the vector $\mathbf{y}$ to have a unit length, the *Maximum A Posteri* estimation of $\mathbf{y}$ is equivalent to maximize the following criterion

$$\mathcal{J} = \frac{\frac{1}{Z} e^{\mathbf{y}^T \mathbf{R}^{-1/2} \tilde{\mathbf{J}} \mathbf{R}^{-1/2} \mathbf{y} + \sum_{i=1}^{l} \tilde{\theta}_i y_i}}{\prod_{i=1}^{l} \left(e^{y_i} + e^{-y_i}\right)},$$

whose numerator has exactly the same form as Eq.(6). However, as the denominator is relevant to $\mathbf{y}$, the maximization of $\mathcal{J}$ becomes very complicated.

So we can see that, our method has a very similar expression as *Bayesian discriminative semi-supervised learning* methods. Although our method is based on statistical physics, while the Bayesian methods are based on Bayesian theory, they are very closely in spirit. And our method makes an approximation that works which omits some of the complications of the *Bayesian* approach. Moreover, the computational complexity for *Naïve Mean Field* approach is $O(N^2)$, while for *Bayesian* methods it is usually $O(N^3)$.

Table 4: Basic information of the benchmark datasets.

Dataset	Classes	Dimension	Size	Type
g241d	2	241	1500	arti.
Digit1	2	241	1500	arti.
USPS	2	241	1500	imba.
COIL	6	241	1500	
BCI	2	117	400	
Text	2	11,960	1500	spar.

6 Experiments

We apply our algorithm to several standard semi-supervised learning datasets[4]. Table 4 provides us some basic information about these datasets.

In Table 4, the first column correspond to the name of the datasets, the second column are the number of classes contained in the dataset, the third and fourth column represent the dimensionality and size of the datasets, and the last column show some properties of the datasets. For the last column, "arti" means *artificial* dataset, "imba" means *imbalanced* dataset, "spar." represents *sparse* dataset. For detailed description of these datasets, one can refer to [5].

For each dataset, there are 100 labeled points, and 12 random splits are performed to partition the dataset into labeled points and remaining unlabeled points. It is ensured that each split contains at least one point of each class. For comparison, we also provide the classification results of 7 semi-supervised learning methods. The *Nearest Neighbor* (*1-NN*) and *Linear SVM* are used as the base line algorithms. The detailed configuration of other semi-supervised learning algorithms are the same as described in [5].

There are two hyperparameters in our method, namely the temperature T and scale σ. In our experiments, they are tuned by a exhaustive search over the grid $2^{[-5:0.5:5]} \times 2^{[-5:0.5:5]}$, where 0.5 is the granularity of the grid. Figure 2 shows the plots of the average test error vs. T and σ, from which we can see that for some of the datasets (*Digit1*), the final classification results are not sensitive to the choice of T and σ, for some

[4] Available at http://www.kyb.tuebingen.mpg.de/ssl-book /benchmarks.html

Table 3: Average test errors (%) with 100 labeled training points

	g241d	**digit1**	**USPS**	**COIL**	**BCI**	**Text**
1-NN	42.45	3.89	5.81	17.35	48.67	30.11
SVM [21]	24.64	5.53	9.75	22.93	34.31	26.45
LLE+1-NN [17]	38.20	2.83	6.50	28.71	47.89	32.83
GRF+CMN [26]	37.49	3.15	6.36	10.03	46.22	25.71
LLGC [23]	28.20	2.77	4.68	9.61	47.67	24.00
TSVM [11]	22.42	6.15	9.77	25.80	**33.25**	24.52
Cluster Kernel [6]	**4.95**	3.79	9.68	21.99	35.17	24.38
LDS [7]	23.74	3.46	4.96	13.72	43.97	**23.15**
Our method	20.41	**1.83**	**4.47**	**8.32**	33.40	24.25

of the datasets (*Text*), the optimal (T, σ) only lies in a small range, and for most of the datasets, the optimal (T, σ) lies in a relatively large area, which makes the parameter tuning procedure easier. We think the different properties of the (T, σ) is because of the different structures of the datasets, and we are currently working on an automatic way to self-tune those parameters.

Table 3 reports the average test errors for various methods, which shows that our mean field approach performs the best on *digit1*, *USPS* and *COIL* datasets, and can produce comparable results on the *BCI* and *Text* dataset. However, as there is no free lunch, it performs fairly poor on the *g241d* dataset. This is probably because the complex structure of the dataset, in which the data from the two classes are heavily overlapped and confused, but our method tends to make the neighboring points have same labels, thus it may confused in this case.

7 Conclusions and Future Works

In this paper, we propose a novel scheme for semi-supervised learning which is based on statistical physics. Unlike the traditional Bayesian methods which aims at doing a maximum a posteri estimation for the labels, our method treats the data labels as a disordered system and the true labels can be regarded as the states of such a system in equilibrium. Many experimental results are presented to show the effectiveness of our method.

In the future, we will focus on two issues of our algorithm: (1) *Induction*, as induction can be viewed as a natural extension of the *cavity method* introduced in section 3, thus deriving an induction approach seems to be a straight forward thing. However, the inclusion of a new spin will affect the state of the original spin system, thus how to tackle such effects is still a hard problem; (2) The *automatic* learning of the hyperparameters in our algorithm; (3) *Acceleration*, the *Naïve Mean Field* approach introduced in section 3 has the computational complexity of $O(n^2)$, which prohibits the usage of our method to large scale datasets, thus deriving effective accelerating methods is also an important direction in our future works.

Acknowledgements

The authors would like to thank the constructive comments of the anonymous reviewers. The work of Fei Wang, Shijun Wang and Changshui Zhang is supported by the China Natural Science Foundation No.60675009.

References

[1] Blatt, M., Wiseman, S., Domany, E. Data Clustering Using a Model Granular Magnet. *Neural Computation* 9, 1805-1842. 1997.

[2] Belkin, M., Matveeva, I., Niyogi, P. Regularization and Semi-supervised Learning on Large Graphs. *COLT*, 2004.

[3] Belkin, M. and Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15: 1373-1396, 2003.

[4] Belkin, M. and Niyogi, P. Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning*, 56: 209-239, 2004.

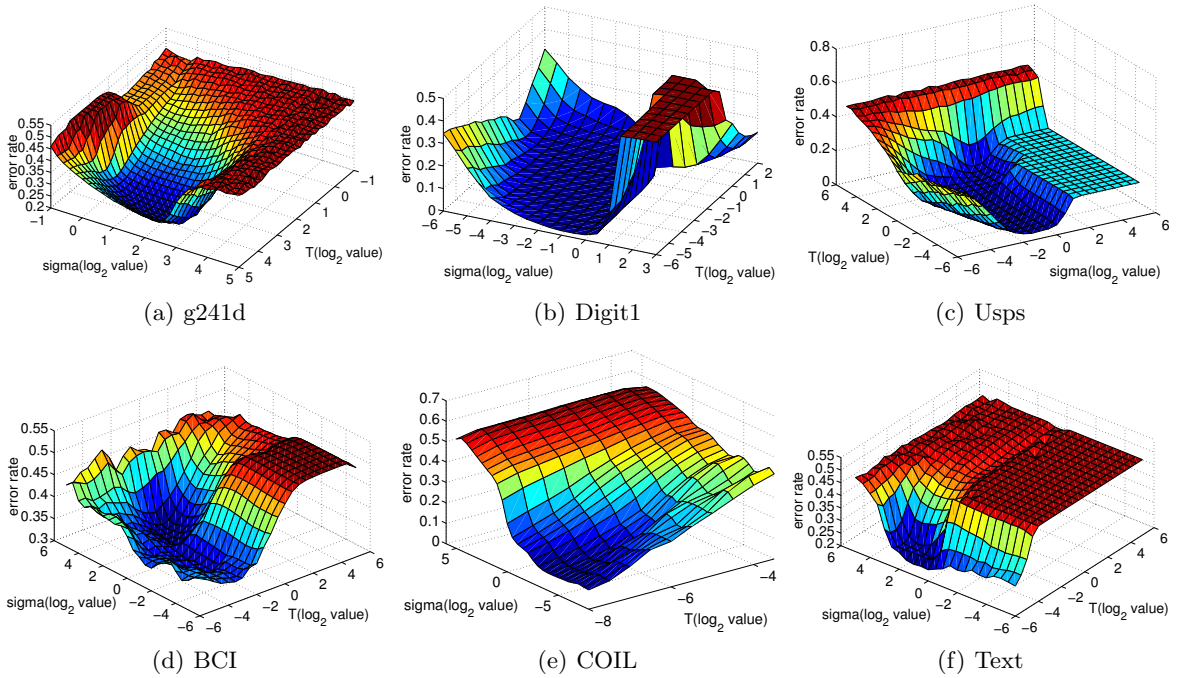
[5] Chapelle, O., Schölkopf, B. and Zien, A. (eds.): *Semi-Supervised Learning*. MIT Press: Cambridge, MA. 2006.

[6] Chapelle, O., Weston, J. and Schölkopf, B. Cluster Kernels for Semi-Supervised Learning. *NIPS* 15, 2003.

[7] Chapelle, O., Zien, A. Semi-Supervised Classification by Low Density Separation. In *AISTATS*. 2005.

[8] Delalleu, O., Bengio, Y. and Le Roux, N. Non-Parametric Function Induction in Semi-Supervised Learning. In *AISTATS*. 2005.

[9] Fischer, B., Roth, V., and Buhmann, J. M. Clustering with the connectivity kernel. In *NIPS*, 16. 2004.



(a) g241d (b) Digit1 (c) Usps

(d) BCI (e) COIL (f) Text

Figure 2: Average error rate vs. temperature T and scale σ. In all the figures, the z-axis represents the average test error, the x-axis is the $\log_2$ value of the temperature T, and the y-axis is the log_2 value of the scale σ.

[10] Ising, E. Beitrag zur theorie der ferromagnetismus. *Z. Phys.* 31, 253-258. 1925.

[11] Joachims, T. Transductive Inference for Text Classification using Support Vector Machines. *ICML*, 1999.

[12] Kapoor, A., Qi, Y, Ahn, H., Picard, R. W. Hyperparameter and Kernel Learning for Graph Based Semi-Supervised Classification. In *NIPS* 18, 2006.

[13] Lal, T. N., Schröder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., Schölkopf, B. Support Vector Channel Selection in BCI. *IEEE Transactions on Biomedical Engineering*, 51(6):1003C1010, 2004.

[14] Mézard, M., Parsi, G., Virasoro, M. A. *Spin Class Theory and Beyond*. Lecture Notes in Physics, 9. World Scientific. 1987.

[15] Opper, M., Winther, O. From Naïve Mean Fields to TAP Equations. In *Advanced Mean Field Methods - Theory and Practice*, eds. M. Opper and D. Saad, 85-98, MIT Press. 2001.

[16] Parisi, G. *Statistical Field Theory*. Redwood City, Calif. : Addison-Wesley Pub. Co. 1988.

[17] Roweis, S. T., Saul, L. K. Noninear Dimensionality Reduction by Locally Linear Embedding. *Science*: vol. 290, 2323-2326, 2000.

[18] Song, Y., Zhang, C., Lee, J., Wang, F. A Discriminative Method For Semi-Automated Tumorous Tissues Segmentation of MR Brain Images. In *MMBIA*, 2006.

[19] Thouless, D. J., Anderson, P. W., Palmer, R. G. Solution of a 'Solvable Model of a Spin Glass'. *Phil. Mag.* 35, 593. 1977.

[20] Tong, S., Koller, D. Restricted bayes optimal classifiers. In *AAAI* 17, pages 658-664, 2000.

[21] Vapnik, V. N. *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1995.

[22] Wang, F., Zhang, C. Label Propagation Through Linear Neighborhoods. *ICML* 23, 2006.

[23] Zhou, D., Bousquet, O., Lal, T. N. Weston, J., & Schölkopf, B. Learning with Local and Global Consistency. In *NIPS* 16, 2004.

[24] Zhou, D., Schölkopf, B. Learning from Labeled and Unlabeled Data Using Random Walks. *DAGM* 26, 2004.

[25] Zhu, X. Semi-Supervied Learning Literature Survey. *Computer Sciences Technical Report* 1530, University of Wisconsin-Madison, 2006.

[26] Zhu, X., Ghahramani, Z., and Lafferty, Z. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML* 20, 2003.

[27] Zhu, X., Lafferty, J., Ghahramani, Z. Semi-Supervised Learning: From Gaussian Fields to Gaussian Process. *Computer Science Technical Report*, Carnegie Mellon University, CMU-CS-03-175. 2003.