Restricted Boltzmann Machine for Classification with Hierarchical Correlated Prior

Gang Chen and Sargur H. Srihari

June 16, 2014

Abstract

Restricted Boltzmann machines (RBM) and its variants have become hot research topics recently, and widely applied to many classification problems, such as character recognition and document categorization. Often, classification RBM ignores the interclass relationship or prior knowledge of sharing information among classes. In this paper, we are interested in RBM with the hierarchical prior over classes. We assume parameters for nearby nodes are correlated in the hierarchical tree, and further the parameters at each node of the tree be orthogonal to those at its ancestors. We propose a hierarchical correlated RBM for classification problem, which generalizes the classification RBM with sharing information among different classes. In order to reduce the redundancy between node parameters in the hierarchy, we also introduce orthogonal restrictions to our objective function. We test our method on challenge datasets, and show promising results compared to competitive baselines.

1 Introduction

Restricted Boltzmann machines (RBM) \cite{11} have attracted significant attention recently on many machine learning problems, such as dimension reduction \cite{10}, text categorization \cite{16}, collaborative filtering \cite{20} and object recognition \cite{14}. A recent survey \cite{1} also shows the advantages of RBMs for classification problems by exploiting prior knowledge about the world around us. Unfortunately, traditional RBM treats the category structure as flat and little work has been done to explore the interclass relationship. We are interested in RBM for classification problems with prior assumption, where the sets of labels are expressed in a hierarchy. The purpose of this paper is to answer whether we can leverage the hierarchical structure over categories to improve the classification accuracy. Organizing different classes in a hierarchy \cite{13, 7, 18, 22, 8, 2, 4} is an efficient and effective way for knowledge representation and categorization. The top level of the taxonomy hierarchies starts with a general or abstract description of common properties for all objects, while the low levers add more specific characteristics. For example, WordNet \cite{7} and ImageNet \cite{5} use this semantic hierarchy to model human psycholinguistic knowledge and object taxonomy respectively, refer to Fig. \ref{fig:example} for more information.

Much work has been done to exploit this hierarchical relationship for classification problems, such
as document categorization [13, 18, 22, 2] and object recognition [17]. One straightforward way to exploit the hierarchy is to classify each node recursively, by choosing the label of which the associated coefficient parameter has the largest output score/probability among its siblings till to a leaf node. For example, the hierarchical classification method with orthogonal transfer proposed by Xiao et al. [23] used the recursive top-down classification strategy. Another example is the nested multinomial logit model [21]. However, the fatal weakness of this strategy for hierarchical classification is that errors will propagate from parent to its children, if any misclassification happened in the parent level.

Another popular methodology for hierarchical classification prefers to use the sum of parameters along the shortest path for classifying cases ended at leaf nodes, such as the hierarchical maximum margin SVM in [2, 4] and the correlated multinomial logit model (corrMNL) in [21]. The basis idea is to decompose classification parameter for each class into contributions from all nodes along the paths from the root of the hierarchy to the leaf associated to this class.

Although the classification restricted Boltzmann machine has yielded promising results [15, 16], it does not exploit the hierarchical prior over label sets to improve hierarchical classification accuracy. In this paper, we generalize RBM with hierarchical prior for classification problems. Basically, we divide the classification RBM into traditional RBM for representation learning and multinomial logit model for classification, see Fig. 2(a) for intuitive understanding. For the traditional RBM (red in Fig. 2(a)), we can extend it into deep belief network (DBN), while for the multinomial logit model (green in Fig. 2(a)), we can incorporate the interclass relationship to it. In this work, we focus on the hierarchical prior over the classification RBM, and we take a similar strategy as corrMNL, that means we use sums of parameters along paths from root to a specific leaf in the tree as model parameters for hierarchical classification. However, we consider it in a rather different way from the previous work. Our contributions are: (1) we introduce the hierarchical semantic prior over labels into restricted Boltzmann machine; (2) we add orthogonal constraints over adjacent layers in the hierarchy, which makes our model more robust for classification problems. We test our method in the experiments, and show comparative results over competitive baselines.

Figure 1: A hierarchical structure example from WordNet [17]. All internal nodes are depicted with white background, while leafs are shown in gray in the hierarchy. This can easily express “cat” and “dog” as carnivore, while “cow” and “sheep” as bovid, and all of them are placental animals.
Figure 2: (a) It is the classification restricted Boltzmann machine, which integrates restricted Boltzmann machine and logistic regression model; the left red dash area is restricted Boltzmann machine for dimension reduction, while the green region shows the logistic regression model for multi-class problems. (b) A hierarchical example for explanation, in which all internal nodes are depicted with white background, while leafs/classes are shown in gray in the hierarchy. The parameters for each classes are presented as a sum of parameters along its ancestors at different level of hierarchy. For example, the coefficient parameter of class 1 is $A_{12} + A_{21}$.

2 Classification restricted Boltzmann machine with hierarchical correlated prior

We will revisit the classification RBM, then we will introduce our model. Throughout the paper, matrix variables are denoted with bold uppercases, and vector quantities are written in bold lowercase. For matrix $W$, we indicate its $i$-th row and $j$-th column element as $W_{ij}$, its $i$-th row vector $W_i$, and $j$-th column vector $W_j$. For different matrixes, we use different subscripts to discern them. For example, $A_{12}$ and $A_{21}$ are different matrixes, which are indicated by different subscripts.

2.1 Classification Restricted Boltzmann Machine

Restricted Boltzmann Machines (RBM) [10] are a particular form of Markov random field (undirected generative model), which are constructed with hidden nodes and visible nodes, and each connection in an RBM must link between a visible node and a hidden node (a bipartite graph, neither connection among visible nodes nor connection among hidden nodes). The classification RBM was first proposed in [12] and was further developed in [15, 16] with discriminative training model. Basically, It [16] generalizes traditional RBM [10] which is undirected graphic model with a layer of hidden variables to model a distribution over visible variables, into the classification RBM with target class observations. To better understand this work, we think the classification RBM is composed of traditional RBM and multinomial logistic regression classifier, see Fig. 2(a) for explanation.

Denote $X \in \mathbb{R}^d$ be an instance domain and $Y$ be a set of labels. Assume that we have a training set $D = \{(x_i, y_i)\}$, comprising for the $i$-th pair: an input vector $x_i \in X$ and a target class $y_i \in Y$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, ..., K\}$. An RBM with $n$ hidden units is a parametric model of the joint distribution between a layer of hidden variables $h = (h_1, ..., h_n)$ and the observations $x = (x_1, ..., x_d)$
and \( y \). The classification RBM joint likelihood takes the form:

\[
p(y, x, h) \propto e^{-E(y, x, h)}
\]

where the energy function is

\[
E(y, x, h) = -h^T W x - b^T x - c^T h - d^T y - h^T U y
\]

with parameters \( \Theta = \{W, b, c, d, U\} \) and \( y = (1_{y=i})_{i=1}^K \) for \( K \) classes, where matrix \( W \in \mathbb{R}^{n \times d} \), and \( U \in \mathbb{R}^{n \times K} \). Further we can compute the following conditional likelihood:

\[
p(h|y, x) = \prod_j p(h_j|y, x) \quad (3a)
\]

\[
p(x|h) = \prod_i p(x_i|h) \quad (3b)
\]

\[
p(x_i = 1|h) = \text{logistic}(b_i + \sum_j W_{ij} h_j) \quad (3c)
\]

\[
p(h_i = 1|v) = \text{logistic}(c_i + \sum_j W_{ji} v_j) \quad (3d)
\]

\[
p(y|h) = \frac{e^{d_y + \sum_j U_{yj} h_j}}{\sum_{y'} e^{d_{y'} + \sum_j U_{y'j} h_j}} \quad (3e)
\]

where \( \text{logistic}(x) = 1/(1 + e^{-x}) \) in Eq. (3c). The prediction \( p(y|h) \) given hidden variables \( h \) in Eq. (3c) is the multinomial logit model (a.k.a multiclass logistic regression or softmax function). In other words, for any new input \( x \), we can encode it into hidden space, and then use \( p(y|h) \) for its label prediction. The green area in Fig. 2(a) shows the prediction with softmax function.

For classification problem, we need to compute the conditional probability for \( p(y|x) \). As shown in [20], this conditional distribution has explicit formula and can be calculated exactly, by writing it as follows:

\[
p(y|x) = \frac{e^{d_y + \sum_j U_{yj} h_j}}{\sum_{y'} e^{d_{y'} + \sum_j U_{y'j} h_j}} \quad (4)
\]

To learn RBM parameters, we need to optimize the joint likelihood \( p(y, x) \) on training data \( D \). Note that it is intractable to compute \( p(y, x) \), because it needs to model \( p(x) \). Fortunately, Hinton proposed an efficient stochastic descent method, namely contrastive divergence (CD) [11] to maximize the joint likelihood. Thus, we get the following stochastic gradient updates for \( W \) and \( U \) from CD respectively

\[
\frac{\partial \log p(x, y)}{\partial W_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \quad (5a)
\]

\[
\frac{\partial \log p(x, y)}{\partial U_{jk}} = \langle h_j y_k \rangle_{\text{data}} - \langle h_j y_k \rangle_{\text{model}} \quad (5b)
\]

And update \( \Theta \) until convergence with gradient descent

\[
\Theta = \Theta + \eta \frac{\partial \log p(x, y)}{\partial \Theta} \quad (6)
\]

where \( \eta \) is the learning rate for RBM.
2.2 Restricted Boltzmann machine with hierarchical prior

Our model introduces hierarchical prior over label sets for logistic regression classifier in the classification RBM. Note that we divide the classification RBM into two parts: RBM (feature learning) and multinomial logit model (classifier), corresponding to red and green regions shown in Fig. 2(a) respectively. Our model introduces the hierarchical prior over multinomial logit regression classifier, which is vital for classification problems under RBM framework.

Define the hierarchical tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, the number of node $N = |\mathcal{V}|$ and the number of edge $M = |\mathcal{E}|$. Furthermore, we assume all parameters along edges are $A = \{A_1, \ldots, A_m\}$, where $\{A_j\}_{j=1}^m$ describes the parameter for each edge in the hierarchy respectively and $A_j$ has the same size as $U$ in the above subsection 2.1. For any node $\nu$ in the tree, we denote $A(\nu)$ as its direct parent (vertex adjacent to $v$), and $A^{(i)}(\nu)$ to be its $i$-th ancestor of $\nu$. As in [4], we also define the path for each node $\nu \in \mathcal{T}$, define $P(\nu)$ to be the set of nodes along the path from root to $v$.

Now we can define the coefficient parameters for each leaf node $\nu$ as

$$A(\nu) = \sum_{\mu \in P(\nu)} A_\mu$$

(8)

where the classification coefficient for each class in Eq. (8) is decomposed into contributions along paths from root to the leaf associated to that class. For our model, each leaf node is associated to one class, which takes the same methodology as in [20]. Fig. 2(b) is an example with total five classes, where the sums of parameters along the path to the leaf node are coefficient parameters used for classification. In Fig. 2(b), $A_{12}$ and $A_{13}$ are parameters along branches in the first level, and $A_{21}$, $A_{22}$, $A_{31}$, $A_{32}$ and $A_{33}$ are parameters in the second level. For example, the coefficient parameter of class 1 is $A_{12} + A_{21}$ according to Eq. (8); similarly, for class 4, its coefficient parameter is $A_{13} + A_{32}$. For example, we can see class 1 and class 2 sharing the common term $A_{12}$, which can be thought as the prior correlation between the parameters of nearby classes in the hierarchy.

For $K$ classes, we have $U \in \mathbb{R}^{n \times K}$ and $A_j \in \mathbb{R}^{n \times K}$ for $j = \{1, \ldots, m\}$. Thus we can factorize

$$U = VA$$

(9)

where $A = \{A_1, \ldots, A_m\} \in \mathbb{R}^{mn \times K}$ is the concatenation of parameters $\{A_j\}_{j=1}^m$ of all edges in the hierarchy, while $V \in \mathbb{R}^{n \times mn}$ implies the hierarchical prior over labels, refer Eq. (8) for construction of the correlated matrix $V$. Note that $V$ (just) encodes given hierarchical structures with 0 or 1 and is fixed during training the models. In addition, we introduce orthogonal restrictions just as in [23] to reduce redundancy between adjacent layers. Given a training set $\mathcal{D} = \{(x_i, y_i)\}$, we propose the following objective function:

$$\mathcal{L}(\mathcal{D}; \Theta) = -\sum_{i=1}^{\mid\mathcal{D}\mid} \log p(y_i, x_i) + C \sum_{\nu, \mu \in P(\nu)} \text{trace}(A^T_{\mu}A_{\nu})$$

(10)

where $C$ is the weight to balance the two terms. The first term is from the negative log likelihood as in RBM and the second term forces parameters at children to be orthogonal to those at its ancestor as much as possible.
The differences between our model and RBM lie: (1) hierarchical prior over labels, which can induce correlation between the parameters of nearby nodes in the tree; (2) we have orthogonal regularization which can make our model more robust, and also reduce redundancy in model parameters. For parameters updating, we have the same equations as in the classification RBM, except for $U$ which introduces hierarchical prior and orthogonal restrictions among children-parent pairs.

According to chain rule, we can differentiate $L(D; \Theta)$ r.w.t $A_\nu$ and get the following derivative

$$\frac{L(D; \Theta)}{\partial A_\nu} = -\sum_{i=1}^{\lvert D \rvert} \log p(y_i, x_i) \cdot \frac{\partial U}{\partial A_\nu} + C \sum_{\mu \in P(\nu)} A_{\mu} \tag{11}$$

Note that the derivative of $\sum_{i=1}^{\lvert D \rvert} \log p(y_i, x_i)$ w.r.t $U$ can be computed with Eq. (5b). Thus, we can use Eq. (11) to calculate derivative w.r.t. $A_\nu$, and then update $A_\nu$ with stochastic gradient descent. Given $A_\nu$, we can use Eq. (9) to update $U$.

### 2.3 Algorithm

Note that our model incorporates the hierarchical prior and orthogonal constraints through $U$. In other words, we can update all parameters with CD, except $U$. Because $U$ is the function of $A$, we can compute the derivative of $U$ w.r.t $A$ and update $A$ with gradient descent. After we get $A$, we can calculate $U$, which can be used in the next iteration. We list the pseudo code below and implement it in Matlab.

**Algorithm 1 Learning RBM with hierarchical correlated prior**

Input: training data $D = \{(x_i, y_i)\}$, the number of hidden nodes $n$, learning rate $\eta$, $C$ and maximum epoch $T$

Output: $\Theta = \{W, b, c, d, U\}$

1: Initialize parameters $W, b, c, d, U$;
2: Divide the training data into batches;
3: for $t = 1$ to $T$ do
4:   for each batch do
5:     Use 1-step Gibbs sampling to update the gradient according to Eq. (5);
6:   end for
7:   Update all other parameters except $U$ with CD;
8:   Compute gradient w.r.t. $A_\nu$ according to Eq. (11);
9:   Update $A$ with gradient descent with Eq. (6);
10:  Update $U$ according to Eq. (9);
11: end for
12: Output $W, b, c, d, U$;
13: End

### 3 Experimental Results

We analyze our model with experiments on two classification problems: character recognition and document classification, and compare our results to those from competitive baselines below.
Figure 3: (a) The hierarchical structure prior over label sets from MNIST digital dataset; we use this prior over labels with the purpose to capture similar structure information between different characters. For example, ‘3’ and ‘8’ share some parts, and similar structure information can be found in pairs ‘4’ and ‘7’, as well as ‘1’ and ‘9’. (b) The hierarchical classification RBM (HRBM). HRBM is constructed according the hierarchical prior (left side graph). In order to learn HRBM classifier, we learn a RBM classifier for each node and recursively to the leaves in a top-down manner.

RBM

The RBM for classification was first proposed in [10] and later was further developed in [16]. Its mathematical formula is shown in Eq. (2).

Hierarchical classification RBM with soft assignment (HRBMs)

The hierarchical classification RBM (HRBM) is a nested hierarchical classifier in a top-down way, shown in Fig. 3(b). In the training stage, for each internal node (including root node) in the current level, HRBM will split training data according to its children and learn a classification RBM for multiple classes (decided by the number of its children). In the inference stage, the likelihood for certain classes in the current layer depends both on the output probability of this layer classifier and also the conditional likelihood on the upper levels. For example, the probability to assign label 2 to a given instance in Fig. 3(b) depends on the output probabilities from \( \text{RBM}_1 \), \( \text{RBM}_{21} \), and \( \text{RBM}_{31} \).

HRBMs computes the classification probabilities for each node in each level, until to leaf nodes. For each data instance, its probability belongs to each class is the probability production along path from root to the leaf of that class, and finally we assign the data instance to the label with largest probability.

Hierarchical classification RBM with hard assignment (HRBMh)

The HRBMh learns RBM classifiers in each level and recursively classify data in a top-down way, shown in Fig. 3(b). More specifically, HRBMh is a special case of HRBM, which assigns labels according to output probabilities in each level until to leafs, instead of assigning probability (soft assignment) as HRBMs in each hierarchical level. The difference between HRBMs and HRBMh is that HRBMs assign classification probability to each node, while HRBMh assign labels.

Hidden hierarchical classification RBM (HHRBM)

Hidden hierarchical classification RBM (HHRBM) is similar as the hierarchical classification RBM (HRBM) in a top-down manner. For any current node, HHRBM learns a classification RBM and projects the training data into hidden space for its children (Note that RBM can map
any input instance into its hidden space). Then, all its children recursively learn classification RBMs with projected hidden states as input from its parent node until to leaf level. In a sense, HHRBM works similar to the deep believe network (DBN) in [12]. Hence, the only difference between HHRBM and HRBM is that HRBM computes the classification probability with the visual data as input for all levels, while HHRBM calculates the classification probability with hidden states as input in a top-down manner.

**Multinomial logit model** (MNL)

MNL, a.k.a multiclass logistic regression, has no class correlated hierarchical structure.

**Correlated Multinomial logit regression** (corrMNL)

corrMNL[1] extends MNL with hierarchical prior over classes, refer to [21] for more details.

In all the above baselines, HRBMs, HRBMh, HHRBM and corrMNL leverage the hierarchical prior over label sets for classification, while RBM and MNL have no such prior information available. As for the difference in the number of RBMs used, (H)HRBMs belong to the tow-down classification approaches where multiple RBMs are constructed and each of which is trained to classify training examples into one of its children in a hierarchical tree while our approach maintains only a single RBM.

![Figure 4: Example images from MNIST dataset.](http://www.ics.uci.edu/~babaks/Site/Codes.html)

**Character Recognition** MNIST dataset[2] consists of 28 × 28-size images of handwriting digits from 0 through 9 with a training set of 60,000 examples and a test set of 10,000 examples, and has been widely used to test character recognition methods. A set of examples are shown in Fig. 4. In the experiment, we use Fig. 3(a) as our hierarchical prior over label sets. To test our method and other baselines, we sample 5000 images from the training sets as our training examples and 1000 examples from the testing sets as our testing data. The reason that we use a subset of MNIST is to answer whether the correlation between different classes is valuable for classification problem when the number of training examples for individual classes may be relatively small. In order to make our method comparable to other baselines, we have the same parameter setting for RBM related methods (including RBM, HRBMs, HRBMh and our method). We set the number of hidden states

---

[1]http://www.ics.uci.edu/~babaks/Site/Codes.html

[2]http://yann.lecun.com/exdb/mnist/
Table 1: The experimental comparison on a subset of MNIST dataset, with total 5000 training examples and 1000 testing samples. We compare the performances between our method and the baselines. It demonstrates that our method with hierarchical prior over labels can improve recognition accuracy.

| Datasets | SVM | MNL | corrMNL | HRBM_h | HRBMs | HHRBM | RBM | Ours |
|----------|-----|-----|---------|--------|-------|-------|-----|------|
| MNIST    | 10.8| 10.6| 8.97    | 12.1   | 7.95  | 11.10 | 8.22| 7.91 |

The comparison between our method and the baselines is shown in Table 1. Our method incorporates the hierarchical prior structure over labels, and the experimental results show that our method outperform other RBM related methods, and also demonstrates that the hierarchical prior in our method is helpful to improve the recognition accuracy.

Document Classification

We also evaluated our model on 20 news group dataset for document classification. The 20 news group dataset has 18,846 articles with with 61188 vocabularies, which has been widely used in text categorization and document classification. In the experiment, we tested our model on the version of the 20 news group dataset, in order to make our results comparable to the current state of the art results.

In the experiment, we used the hierarchical prior structure over label shown in Fig. 5 for HHRBM, HRBM_h, HRBMs and our model. As for parameter setting, we use CD-1, and set the number of hidden states \( n = 2000 \), learning rate \( \eta = 0.1 \) and the maximum epoch equals to 100 for RBM related methods. For HHRBM, we set the number of hidden states to be 1000, 500, 200 and 200 respectively for each layer. As for our method, we set \( n = 2000, \eta = 0.01, C = 0.1 \) and maximum epoch 200.

The results of different methods are shown in Table 2. Once again, our model outperforms the other RBM models, also get better results than SVM and neural network classifiers. HRBMs and corrMNL has bad performance in this dataset. The reason we guess is that HRBMs calculates the classification probability for each class by multiplying the output probabilities along the path from root to the leaf associated to that class. Thus, HRBMs will prefer the high level class for unbalanced hierarchical structure. Note that the hierarchical tree in Fig. 5 is unbalanced structure. For HRBMs, ‘alt.Atheism’, ‘misc.forsale’ and ‘soc.religon.christian’ will have higher probability to be labeled compared to leaves (or classes) in the level 4. corrMNL may have the same problem as HRBMs. Another reason for the low performance is that corrMNL does not consider the parameter redundancy problem between adjacent layers as in our model.
Figure 5: The hierarchical structure from 20 news group dataset. The root (or the first level) cover documents from all categories, while the leaf level indicates labels where documents attached to.

| Model            | Error rate (%) |
|------------------|----------------|
| RBM              | 24.9           |
| DRBM             | 27.6           |
| RBM + NNet       | 26.8           |
| HDRBM            | 23.8           |
| HRBMh (η = 0.1, n = 2000) | 30.6         |
| HRBMs (η = 0.1, n = 2000) | 63.7         |
| HHRBM (η = 0.1, n = 1000, 500, 200 and 200) | 32.0         |
| Ours (η = 0.01, n = 2000 and C = 0) | 30.4         |
| Ours (η = 0.01, n = 2000 and C = 0.1) | 23.76        |
| MNL              | 30.8           |
| corrMNL          | 79.3           |
| SVM              | 32.8           |
| NNet             | 28.2           |

Table 2: The experimental comparison on 20 news group dataset. We compare the performances between our method and other RBM models. It demonstrates that our method with hierarchical prior over labels can improve recognition accuracy.

We also evaluate how the regularization term influence the performance. We set C = 0 to remove the orthogonal restriction, and get accuracy 30.4% in Table 2. It demonstrates that it is useful to add orthogonal restriction to the correlated hierarchical prior.

4 Related work

The hierarchical structure is organized according to the similarity of classes. Two classes are considered similar if it is difficult to distinguish one from the other on the basis of their representation. The similarity of classes increases as we descend the hierarchy. Thus, the hierarchical prior over categories provides semantic meaning and valuable information among different classes; and thus to some extent it can assist classification problems in hand. There has much work extensively been done in the past years to exploit hierarchical prior over labels for classification problem, such as document categorization and object recognition.
Two most popular approaches to leverage hierarchical prior can be categorized below. The first approach classifies each node recursively, by choosing the label of which the associated vector has the largest output score among its siblings till to a leaf node. An variant way is to compute the conditional probability for each class at each level, and then multiply these probabilities along every branch to compute the final assignment probability for each class. Xiao et al. introduced a hierarchical classification method with orthogonal transfer [23], which requires the parameters of children nodes are orthogonal to those of its parents as much as possible. Another example is the nested multinomial logit model [21], in which the nested classification model for each node is statistically independent, conditioned on its parent in the upper levels. One weakness of this strategy for hierarchical classification is that errors will propagate from parents to children, if any misclassification happened in the top level. The other methodology for hierarchical classification prefers to use the sum of parameters along the tree for classifying cases ended at leaf nodes. Cai and Hoffmann [2] proposed a hierarchical larger margin multi-class SVM with tree-induced loss functions. Similarly, Dekel et al. in [4] improved [2] into an online version for hierarchical classification. Recently, Shahbaba et al. proposed a correlated multinomial logit mode (corrMNL) [21], whose regression coefficients for each leaf node are represented by the sum of parameters on all the branches leading to that class.

Apart from the two approaches mentioned above, there are also other methods proposed in the past. Dumais and Chen trained different classifiers kind of layer by layer by exploring the hierarchical structure [6]. Cesa-Bianchi et al. combined Bayesian inference with the probabilities output from SVM classifiers in [3] for hierarchical classification. Similarly, Gopal et al. [9] used Bayesian approach (with variational inference) with hierarchical prior for classification problems.

To the best of our knowledge, no work until now has incorporated hierarchical prior into RBM framework. In this paper, we propose a classification restricted Boltzmann machine with hierarchical correlated prior. Basically, we divide the classification RBM into traditional RBM (feature learning or representation) and multi-class logistic regression, and introduce the hierarchical prior over categories into the logistic prediction part in RBM. Thus, our model can capitalize the valuable information endowed with the taxonomy. Moreover, we also force orthogonal constraints between parameters along children-parent pairs in the hierarchy.

5 Conclusion

We consider restricted Boltzmann machines (RBM) for classification problems, with prior knowledge of sharing information among classes in a hierarchy. Basically, our model decompose classification RBM into traditional RBM and multi-class logistic regression model, and then introduce hierarchical prior over multi-class logistic model. In order to reduce the redundancy between node parameters, we also introduce orthogonal restriction in our objective function. We test our method on challenge datasets, and show promising results compared to benchmarks. In the further work, we will generalize our model with deep belief network or sparsity. Another possible topic is to apply our model to multi-label problems.
References

[1] Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. TPAMI (2012)

[2] Cai, L., Hofmann, T.: Hierarchical document categorization with support vector machines. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. pp. 78–87. CIKM ’04, ACM, New York, NY, USA (2004)

[3] Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Hierarchical classification: Combining bayes with svm. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 177–184. ICML ’06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1143844.1143867

[4] Dekel, O., Keshet, J., Singer, Y.: Large margin hierarchical classification. In: Proceedings of the Twenty-first International Conference on Machine Learning. pp. 27–34. ICML ’04, ACM, New York, NY, USA (2004)

[5] Deng, J., Dong, W., Socher, R., jia Li, L., Li, K., Fei-fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)

[6] Dumais, S., Chen, H.: Hierarchical classification of web content. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 256–263. SIGIR ’00, ACM, New York, NY, USA (2000), http://doi.acm.org/10.1145/345508.345593

[7] Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. The MIT Press (May 1998)

[8] Goodman, J.: Classes for fast maximum entropy training. ICASSP (2001)

[9] Gopal, S., Yang, Y., Bai, B., Niculescu-Mizil, A.: Bayesian models for large-scale hierarchical classification. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS. pp. 2420–2428 (2012)

[10] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (Jul 2006)

[11] Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. 14(8), 1771–1800 (Aug 2002)

[12] Hinton, G.E.: Learning multiple layers of representation. Trends in Cognitive Sciences 11, 428–434 (2007)

[13] Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: Proceedings of the Fourteenth International Conference on Machine Learning. pp. 170–178. ICML ’97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)

[14] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
[15] Larochelle, H., Bengio, Y.: Classification using discriminative restricted boltzmann machines. In: Proceedings of the 25th International Conference on Machine Learning. pp. 536–543. ICML '08, ACM, New York, NY, USA (2008)

[16] Larochelle, H., Mandel, M., Pascanu, R., Bengio, Y.: Learning algorithms for the classification restricted boltzmann machine. J. Mach. Learn. Res. 13(1), 643–669 (Mar 2012), http://dl.acm.org/citation.cfm?id=2503308.2188407

[17] Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR. IEEE Computer Society (2007)

[18] McCallum, A., Rosenfeld, R., Mitchell, T.M., Ng, A.Y.: Improving text classification by shrinkage in a hierarchy of classes. In: Proceedings of the Fifteenth International Conference on Machine Learning. pp. 359–367. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998)

[19] Rohit, B., Ioannis, P., Eric, G., Massih-Reza, A.: On flat versus hierarchical classification in large-scale taxonomies. In: Neural Information Processing Systems (NIPS) (2013)

[20] Salakhutdinov, R., Mnih, A., Hinton, G.: Restricted boltzmann machines for collaborative filtering. In: Proceedings of the 24th International Conference on Machine Learning. pp. 791–798. ICML '07, ACM, New York, NY, USA (2007)

[21] Shahbaba, B., Neal, R.M.: Improving classification when a class hierarchy is available using a hierarchy-based prior. Bayesian Analysis 2(1), 221–237 (2007)

[22] Weigend, A.S., Wiener, E.D., Pedersen, J.O.: Exploiting hierarchy in text categorization. Inf. Retr. 1(3), 193–216 (1999)

[23] Xiao, L., Zhou, D., Wu, M.: Hierarchical classification via orthogonal transfer. In: ICML. pp. 801–808 (2011)