Learning robust features from underwater ship-radiated noise with mutual information group sparse DBN

Sheng SHEN1; Honghui Y ANG1; Zhen HAN1; Junjun SHI1; Jinyu XIONG2; Xiaoyong ZHANG2
1 School of Marine Science and Technology, Northwestern Polytechnical University, China
2 Science and Technology on Blind Signals Processing Laboratory, China

ABSTRACT
Classifying Underwater Acoustic Target(UAT) based on ship-radiated noise datasets is a small-sample-size classification problem. Generalization ability of features depending on prior knowledge and traditional signal processing is seriously restricted for lack of labeled data. Deep Belief Networks (DBN) can make use of abundant unlabeled data and extract features spontaneously, but the resulting features exhibit statistical dependencies. In this paper, we propose a feature learning model called mutual information group sparse DBN (MIGS-DBN) for ship-radiated noise. Firstly, a standard RBM is pre-trained in an unsupervised learning way, then the hidden units are grouped adaptively according to mutual information and the activation level are punished. Secondly, the improved RBMs are stacked by performing a greedy layer-wise training phase, followed by the discriminative fine-tuning to achieve the sparse representation of multilayer features. Recognition experiments on real-world ship-radiated noise show that, features learned by the proposed model can reduce the redundancy and improve classification performance significantly.

Keywords: Ship-radiated noise, Machine Learning, DBN

1. INTRODUCTION
Identification of underwater acoustic signal is an important field of pattern recognition. In order to improve the accuracy of the underwater target recognition, the radiate noise of the original signal is analyzed to extract features such as time-domain waveform features, wavelet features, and auditory features (1, 2, 3). Waveform features are extracted according to probability distribution of wavelength, spectral analysis with zero crossing and the probability distribution of amplitude (1). A wavelet is a mathematical function used to divide a given function or continuous-time signal into different scale components (2). Auditory spectrum features are extracted according to critical bands and scale model of loudness to simulate the psychophysics of auditory system (3). Features above need to be extracted based on expert knowledge and prior knowledge to design feature extraction algorithm or adjust parameters. However, the ship radiated noise results from the synergetic action of various factors, such as machinery noise, propeller noise and hydrodynamic noise (4). The mechanism of these noises is not identical. The same target of different motions has different features. And the marine environment is relatively complex. Different depth has different environmental noise source. Different frequency has different noise characteristics. In addition, the multipath effect and Doppler frequency shift can also cause the decline of the signal and waveform distortion (5).

Traditional features of underwater acoustic target are susceptible to environment and the state of target. With the increase of the complexity of the task, generalization ability of traditional features is seriously restricted for lack of labeled data. However, RBM can be trained using plenty of unlabeled data and they can learn stochastic features for modeling the higher-order statistical structure of a dataset (6). But features learned by RBM always have high statistical correlation due to the learning algorithm.

We describe a MIGS-DBN that uses the proposed mutual information group sparse RBM (MIGS-RBM) to construct higher-level features using the unlabeled data. Features learned by the proposed model form a succinct input representation and significantly improve classification performance.

This paper is organized as follows. The theory of RBM and its problem are presented in Section 2. The MIGS-RBM and MIGS-DBN algorithms we proposed are described in Section 3. Experimental results and discussion are shown in Section 4. Conclusion is summarized in Section 5.

2. RBM
2.1 RBM
RBM is a bipartite graph in which visible units \( v = (v_1, v_2, ..., v_v) \) that represent observations are connected to hidden units \( h = (h_1, h_2, ..., h_h) \) that learn to represent features using undirected weighted
connections. There are no visible-visible or hidden-hidden connections. To deal with real-valued
ship-radiated noise, we use a Gaussian-Bernoulli RBM in which the hidden units are binary but the input
units are linear with Gaussian noise. The weights and biases of the individual units define a probability
distribution over the joint states of the visible and hidden units via an energy function (7):
\[
E(v, h | \theta) = \frac{1}{2} \sum_{i=1}^{n} (v_i - a_i)^2 + \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n} \sum_{j=1}^{m} v_i W_{ij} h_j
\]
Where \( \theta = \{ W_i, a_i, b_j \} \), and \( W_i \) represents the symmetric interaction term between visible unit \( i \) and
hidden unit \( j \), while \( a_i \) and \( b_j \) are their bias terms. \( n \) and \( m \) are the numbers of visible and hidden units.
The conditional distribution \( P(h_j = 1 | v, \theta) \) is given by:
\[
P(h_j = 1 | v, \theta) = \sigma(b_j + \sum_i v_i W_{ij})
\]
Where \( \sigma(x) = 1 / (1 + e^{-x}) \)
Similarly, the conditional distribution \( P(v_i = 1 | h, \theta) \) is given by:
\[
P(v_i = 1 | h, \theta) = N(a_i + \sum_j W_{ij} h_j, 1)
\]
Where \( N(\mu, \nu) \) is a Gaussian with mean \( \mu \) and variance \( \nu \).
The gradient of the log probability of the training data is:
\[
\frac{\partial \log p(v)}{\partial \theta} = -E_{v,h}[\frac{\partial}{\partial \theta} E(v, h)] + E_{v,h}[\frac{\partial}{\partial \theta} E(v, h)]
\]
Where \( E \) is the expectation.

2.2 RELEVANCY OF HIDDEN UNITS
The goal of learning algorithm is to maximize the likelihood of training data. Given observation data
\( v \) and the reconstructed data \( v' \), the gradient of weights and biases corresponding to hidden unit \( j \) is
\[
\Delta W_{ij} = P(h_j = 1 | v) \cdot v - P(h_j = 1 | v') \cdot v'
\]
\[
\Delta b_j = P(h_j = 1 | v) - P(h_j = 1 | v')
\]
The probability that the network assigns to a training data \( v \) can be raised by adjusting the weights and
biases to lower the energy of that train data \( v \). In other words, the hidden unit \( j \) is learning to represent the
training data. Because the hidden states of different hidden units are conditionally independent, hidden units
will independently learn to represent the training data. Therefore, parts of the hidden units are likely to learn
similar characteristics, behaving as strong statistical correlation. Hidden units which are only rarely active are
typically easier to interpret than those which are active about half of the time. Also, discriminative
performance is sometimes improved by using features that are only rarely active (8).
So we attempt to consider the statistical correlation between hidden units in the process of sparse
procedure to avoid multiple hidden units learning similar characteristics.

3. MIGS-RBM AND MIGS-DBN
Sparse activities of the binary hidden units can be achieved by specifying a “sparse penalty
function”. Sparse methods always assume that features are independent to each other. But the features
learned by RBM often show a strong statistical correlation, and many features often appear in groups
(9). To alleviate this problem, we propose the strategy of grouping the hidden units by mutual
information, and then add the sparse punishment mechanism that uses \( 1/2 \) regularization upon the
activation probabilities of hidden units to capture the local dependencies among hidden units. Finally,
using the greedy training algorithm to build DBN framework based on the RBM present before.

3.1 The Mutual Information Grouping Strategy
Information theory provides intuitive tools to quantify the uncertainty of random hidden units, or
how much information is shared by a few of them. Mutual information between two variables is the
reduction of uncertainty of one variable while the other is known (10, 11, 12).
The dependence of two variables measured by the mutual information is as follow:
\[
I(X; Y) = \text{KL}[p(x, y) || p(x)p(y)]
\]
\( \text{KL}[p \| q] \) is the Kullback-Leibler divergence of two probability functions \( p \) and \( q \) for discretized
random variables.
\[
\text{KL}[p(x) \| q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)}
\]
After a standard RBM is pre-trained in an unsupervised learning way, we group the hidden units
adaptively according to mutual information to restrain the dependencies within these groups.
By calculating the mutual information between \( h \) and other hidden units, we average the hidden units
into groups according to \( \text{Relev} \), along with the weights and biases. The correlation of units is larger in the

5138
same group. 

\[ \text{Relev}_j = \frac{1}{m} \sum_{j=1}^{m} I(h_j; h_j) \]  

\[ \text{(8)} \]

3.2 Sparse Penalty Function

We introduce a mixed-norm $\ell_1/\ell_2$ regularizer on the activation probabilities of hidden units grouped by mutual information. The mixed-norm is given by (9):

\[ \sum_{k=1}^{K} N_k(v) = \sum_{k=1}^{K} \sum_{m \in G_k} P(h_m = 1 | v)^2 \]  

\[ \text{(9)} \]

The hidden layer is divided into $K$ groups. The $k^{th}$ group is denoted by $G_k$. We add the $\ell_1/\ell_2$ regularizer to the log-likelihood of training data.

\[ \max_{\theta} \sum_{l=1}^{L} \log P(v) - \lambda \sum_{k=1}^{K} N_k(v) \]  

\[ \text{(10)} \]

Where $\lambda$ is constant.

By introducing this regularizer, Equation 5 and 6 are changed to the following equation

\[ \Delta w_j = P(h_j = 1 | v) - P(h_j = 1 | v') \cdot v - \lambda \alpha_j \]  

\[ \Delta b_j = P(h_j = 1 | v) - P(h_j = 1 | v') \cdot v - \lambda \beta_j \]  

\[ \text{(11)} \]

\[ \text{(12)} \]

Where

\[ \alpha_j = \frac{\partial}{\partial w_j} \sum_{k=1}^{K} N_k(v) = \frac{1}{N_k(v)} P(h_j = 1 | v)^2 P(h_j = 0 | v) \cdot v \]  

\[ \text{(13)} \]

\[ \beta_j = \frac{\partial}{\partial b_j} \sum_{k=1}^{K} N_k(v) = \frac{1}{N_k(v)} P(h_j = 1 | v)^2 P(h_j = 0 | v) \]  

\[ \text{(14)} \]

Table 1 - Training with $\ell_1/\ell_2$ regularizer

<table>
<thead>
<tr>
<th>Input: training sample $x$; learning rate $\eta$; regularization constant $\lambda$</th>
<th>Output: $W, a, b$</th>
</tr>
</thead>
<tbody>
<tr>
<td># Gibbs sample</td>
<td></td>
</tr>
<tr>
<td>initialization: $v = x$</td>
<td></td>
</tr>
<tr>
<td>sample $h$ from $P(h</td>
<td>v)$</td>
</tr>
<tr>
<td>sample $v'$ from $P(v'</td>
<td>h)$</td>
</tr>
<tr>
<td>sample $h'$ from $P(h'</td>
<td>v')$</td>
</tr>
<tr>
<td># Calculate the regularizer</td>
<td></td>
</tr>
<tr>
<td>For $k = 1: K$</td>
<td></td>
</tr>
<tr>
<td>For $j = 1: m / K$</td>
<td></td>
</tr>
<tr>
<td>$\alpha_j = \frac{1}{N_k(v)} P(h_j = 1</td>
<td>v)^2 P(h_j = 0</td>
</tr>
<tr>
<td>$\beta_j = \frac{1}{N_k(v)} P(h_j = 1</td>
<td>v)^2 P(h_j = 0</td>
</tr>
<tr>
<td>Where $h_j \in G_k$</td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
</tr>
<tr>
<td># Update with regularizer</td>
<td></td>
</tr>
<tr>
<td>$W = W + \eta[P(h</td>
<td>v)v - P(h'</td>
</tr>
<tr>
<td>$a = a + \eta(v - v')$;</td>
<td></td>
</tr>
<tr>
<td>$b = b + \eta[P(h</td>
<td>v) - P(h'</td>
</tr>
</tbody>
</table>

We assume the $j^{th}$ hidden unit belongs to the $k^{th}$ group. This regularization not only encourages hidden units of many groups to be inactive but also makes hidden units within a group compete with each other for modeling observed data, meanwhile keeps the exact and efficient inference in RBMs. By considering statistical dependencies of states of hidden units we may learn a more powerful generative model (13).

3.3 Learning MIGS-DBN

The MIGS-DBN in the paper is motivated by the observation that even many unknown categories of ship radiated noise will contain basic patterns that are similar to those in targets. We can use adequate unlabeled data effectively to initialize the network.
After training an MIGS-RBM on the unlabeled data, the parameters are frozen, and the inferred states of the hidden units can be used as data for training another MIGS-RBM. By repeatedly applying such a procedure, we can learn a MIGS-DBN which has many layers to that represent more complex statistical structure in the data.

The use of features found by the MIGS-DBN to initialize a multilayer of back-propagation network significantly decreases both the time taken for discriminative training and the amount of over fitting.

4. EXPERIMENTS AND DISCUSSION

Work conducted in this paper was performed on the real-world underwater acoustic noise recorded using passive sonar equipment. The unlabeled data include marine background noise, unconfirmed ship-radiated noise and so on. The labeled data are divided into 2 categories on behalf of the different types of ship. We split the labeled dataset in the following way: 70% for training and 30% for testing. We divided the audio into short frames. For each of these frames, we calculated the discrete Fourier transform.

The MIGS-DBN was first pre-trained with the unlabeled dataset in an unsupervised manner. We then proceeded to the supervised fine-tuning using the labeled training set.

4.1 Visualization

4.1.1. Sparseness of Features

RBM and MIGS-RBM contains 50 hidden units respectively that are used to observe the activation probability of the hidden layer on a sample. As shown in Figure 1, compared with the standard RBM, the probability of hidden layer of MIGS-RBM is more sparse.

In order to quantify the sparse degree of hidden layer, we introduce the sparseness (14) defined as follow:

$$\text{sparseness}(v) = \frac{\sqrt{D} - \frac{1}{D} \sum_{i=1}^{D} v_i^2}{\sqrt{D} - 1}$$

(15)

The value stands at [0, 1], while the value is greater, the sample is more sparse. After a standard RBM has trained, the sparseness of hidden units stands at [0.002, 0.05] and the mean value is 0.034. The sparseness of MIGS-RBM stands at [0.08, 0.43], and mean value is 0.25. Sparse degree of MIGS-RBM was obviously higher than that of standard RBM. Selecting 1000 samples randomly, the sparseness of standard RBM and MIGS-RBM are shown in Figure 2, for almost all samples, the sparse degree of MIGS-RBM is much better than that of standard RBM.
4.1.2 Redundancy of Features

Greater mutual information represents more redundancy. The mutual information of any two hidden units of standard RBM and MIGS-RBM is shown in Figure 3(a) and (b) respectively.

\[ \text{Relev} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} I(h_i; h_j) \]  \hspace{1cm} (16)

4.1.3 Clustering

In Figure 4 we have plotted a 2-dimensional projection of the representations by using the t-SNE algorithm (15). The clustering of the activations of the hidden layers indicates that the features in MIGS-RBM will benefit the accuracy of the classifiers.
4.2 Classification

In order to verify the validity of sparse penalty factor for learning features, we conduct a comparison experiment from three aspects: 1) sparse degree of features; 2) redundancy of features; 3) recognition performance.

Table 2 shows the performance of RBM and MIGS-RBM on ship-radiated noise using nets with one hidden layer with 20, 50, 100, 200, 300, 400 and 500 hidden units. The redundancy and sparseness defined above are calculate respectively. With the increase of the number of the hidden units, the redundancy decrease gradually and the sparse degree increase gradually. The average sparse degree of MIGS-RBM is greater than that of RBM showing that the method can restrain the activation of the hidden unit, therefore, the sparse degree is improved. In addition, the redundancy of hidden units learned by MIGS-RBM is lower than that learned by RBM, showing that the proposed method can reduce the redundancy between the learned features. The last two columns compare recognition accuracy of RBM and MIGS-RBM. Note that the MIGS-RBM outperforms RBM in most cases.

<table>
<thead>
<tr>
<th>Hidden Unit</th>
<th>RBM Redundancy</th>
<th>RBM Sparseness</th>
<th>RBM Accuracy</th>
<th>MIGS-RBM Redundancy</th>
<th>MIGS-RBM Sparseness</th>
<th>MIGS-RBM Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.7038</td>
<td>0.5426</td>
<td>0.0001</td>
<td>0.0069</td>
<td>0.723</td>
<td>0.719</td>
</tr>
<tr>
<td>50</td>
<td>0.5679</td>
<td>0.3948</td>
<td>0.0019</td>
<td>0.0964</td>
<td>0.791</td>
<td>0.811</td>
</tr>
<tr>
<td>100</td>
<td>0.3735</td>
<td>0.1697</td>
<td>0.0072</td>
<td>0.2297</td>
<td>0.813</td>
<td>0.829</td>
</tr>
<tr>
<td>200</td>
<td>0.1954</td>
<td>0.1006</td>
<td>0.0202</td>
<td>0.3354</td>
<td>0.824</td>
<td>0.847</td>
</tr>
<tr>
<td>300</td>
<td>0.1152</td>
<td>0.0724</td>
<td>0.0296</td>
<td>0.3602</td>
<td>0.849</td>
<td>0.861</td>
</tr>
<tr>
<td>400</td>
<td>0.008</td>
<td>0.0627</td>
<td>0.0356</td>
<td>0.3630</td>
<td>0.842</td>
<td>0.860</td>
</tr>
<tr>
<td>500</td>
<td>0.0571</td>
<td>0.0455</td>
<td>0.0415</td>
<td>0.3455</td>
<td>0.837</td>
<td>0.857</td>
</tr>
</tbody>
</table>

We learn a MIGS-DBN with 3 hidden layers. As shown in Table 3 the accuracy is better than that of a single layer network shown in Table 2.

<table>
<thead>
<tr>
<th>MIGS-DBN</th>
<th>Hidden units</th>
<th>RBM Redundancy</th>
<th>RBM Sparseness</th>
<th>RBM Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS-RBM1</td>
<td>300</td>
<td>0.0724</td>
<td>0.3629</td>
<td></td>
</tr>
<tr>
<td>MIGS-RBM2</td>
<td>300</td>
<td>0.1260</td>
<td>0.1657</td>
<td>0.867</td>
</tr>
<tr>
<td>MIGS-RBM3</td>
<td>300</td>
<td>0.1860</td>
<td>0.1621</td>
<td></td>
</tr>
</tbody>
</table>

The confusion matrix corresponds to Table 3 is as follow, which shows the two categories can be separated preferably.

<table>
<thead>
<tr>
<th>Target/Output</th>
<th>First class</th>
<th>Second class</th>
</tr>
</thead>
<tbody>
<tr>
<td>First class</td>
<td>0.833</td>
<td>0.099</td>
</tr>
<tr>
<td>Second class</td>
<td>0.167</td>
<td>0.901</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

We present a deep feature learning architecture composed of improved RBM which punish the activation level of the grouped hidden units by a mixed-norm to achieve the sparse representation of multilayer features. By implementing our algorithm on the real world underwater acoustic target data, we show that the proposed algorithm can increase the sparseness and reduce the redundancy of features significantly while maintaining recognition accuracy.

ACKNOWLEDGEMENTS

This work is in part supported by grants from Science and Technology on Blind Signals Processing Laboratory.

REFERENCES


