LEUKEMIA CLASSIFICATION USING DEEP BELIEF NETWORK

Wannipa Sae-Lim, Wiphada Wettayaprasit, Pattara Aiyarak*
Artificial Intelligence Research Laboratory, Department of Computer Science, Faculty of Science
Prince of Songkla University
Songkhla, Thailand
giftzn@hotmail.com, wwettayaprasit@yahoo.com, pattara@aiyarak.net

ABSTRACT
This paper proposes a novel approach for leukemia classification based on the use of Deep Belief Network (DBN). DBN is a feedforward neural network with a deep architecture that consists of a stack of restricted Boltzmann machine (RBM). The study used the benchmark DNA microarray of leukemia data from Kent Ridge Bio-medical Data Set Repository. The classification performance was compared between the proposed method and the traditional neural networks. In conclusion, the DBN outperforms the state-of-the-art learning models such as support vector machine (SVM), k-nearest neighbor (KNN) and Naive Bayes (NB).

KEY WORDS
Deep belief network, restricted Boltzmann machine, neural networks, microarray data, leukemia classification

1. Introduction
One important method to deal with the gene expression microarray data is the classification of the type of tumor [1]. Cancer classification is a main research area in the medical field. Such classification is an essential step in prediction and diagnosis of diseases [2]. Acute leukemia is a type of cancer of the bone marrow characterized by an abnormal increase of immature white blood cells that cannot fight infection. When lymphoid cells are affected, the disease is called acute lymphoblastic leukemia (ALL); however, it is called acute myeloid leukemia (AML), when myeloid cells are affected [3]. The classification of leukemia is a way to differentiate cancer tissues such as ALL and AML for using in medical diagnosis and hence a suitable treatment. The classification accuracy is essentially very important point for cancer treatment. Many machine-learning techniques have been developed to reach highly accurate classification performance when classifying DNA microarray data.

DNA microarray technology is an advanced tool used in molecular biology and biomedicine for measuring the large-scale gene expression profiles under varying biologic tissues and for producing a great amount of genetic data. DNA microarray consists of an arrayed series of spots of DNA, called features [4]. The analysis and understanding of microarray data includes a search for genes that have similar or correlated patterns of expression. Among the various frameworks in which pattern recognition has been traditionally formulated, statistical approach, neural network techniques and methods imported from statistical learning theory are among those that have been applied in microarray data analysis [5]. This technology has been widely used in many fields such as drug screening, agriculture as well as the clinical diagnosis of human diseases. To address the classification of cancer problem, we use Deep Belief Network (DBN) that is one technique in machine learning.

Several machine learning techniques have been previously used in classifying gene expression data including k nearest neighbor, decision tree, multilayer perceptron, support vector machine, boosting and self-organizing map [6]. Moreover, deep architecture such as DBN have used in many tasks and shown good performance in various real-world applications. For example, Larochelle et. al. presented the experiments which indicate that deep architecture can solve the learning problems with many factors of variation and outperform SVMs and single hidden layer feedforward neural networks [7]. In addition, Salakhutdinov and Hinton showed how to use unlabeled data and a DBN to learn a good covariance kernel for a Gaussian process [8]. Horster and Lienhart also investigated deep network for image retrieval on large-scale databases [9]. Mobahi et. al. proposed a learning method for deep architectures that took advantage of sequential data [10]. Furthermore, Salama et. al. applied the DBN with continuous data for clustering and classification [11]. Currently, DBN has been applied to different applications, such as audio classification [12], object recognition [13] and many visual data analysis tasks [14,15].

The discovery to efficient training procedure for deep learning came in 2006 with the algorithm for training deep belief network by Hinton et. al. [16] and stacked auto-encoders, which are based on a similar method, greedy layer-wise unsupervised pre-training followed by supervised fine-tuning by Ranzato et. al. in 2007 [17]. DBN is a probabilistic generative models that contain many layers of hidden layers including a greedy layer-wise learning algorithm. Each layer is pre-trained with an unsupervised learning algorithm, learning a nonlinear transformation of its input that captures the interesting feature as the input data for the next higher layer. The main building block of a DBN is a bipartite undirected graphical model called a restricted Boltzmann Machine (RBM) [18]. An RBM is undirected generative models that use a layer of hidden variables to model a distribution over visible variables. It has a single layer of hidden units, which are not connected to each other, and have undirected symmetrical connections to a layer of visible units. This RBM is called Harmonium RBM [19].
This study used the benchmark DNA microarray of leukemia data from Kent Ridge Bio-medical Data Set Repository [20] which was first adopted into the classification of leukemia by Golub et al. in 1999 [21].

In this paper, we focus on a machine learning approach for identifying cancer types based on DBN and evaluated its performance on DNA microarray data of leukemia. We also compare DBN to the well known classifier of SVM, KNN and NB. The experimental results indicate that DBN provides better performance than SVM, KNN and NB in leukemia classification.

This paper is organized as follows. After introducing the concept in Section 1, the backgrounds of RBM and DBN are shown in Section 2. Section 3 describes the DBN approach for classification of leukemia with training data algorithm. Section 4 presents the experimental setting and experimental results along with the corresponding discussions. The conclusions and the direction for future study of this paper are addressed in Section 5.

2. Deep Belief Network for Leukemia Classification

2.1 Generative vs. Discriminative models

Two main types of probabilistic models are generative and discriminative models. The difference between the two models is the probability distribution. Basically the goal of training is to find the conditional distribution $p(y|x)$ to predict some output $y$ given the value of an input $x$. Discriminative models (such as traditional feedforward neural networks) generate the probability of an output given an input that the result can be used to make predictions of $y$ for a new value of $x$. On the other hand, the generative models (such as DBN) generate the joint probability distribution of the input and the output. The generative model estimate $p(x,y)$ that is possible to obtain either $p(y|x)$ or $p(x|y)$ using Bayes' theorem. Although the classification error rate of generative model is generally greater than discriminative model, the last and only supervised fine-tuning phase in DBN training is sufficient update the weights to minimize the appropriate loss function directly.

2.2 Restricted Boltzmann Machine

A Restricted Boltzmann Machine (RBM) is the core component of the DBN with no connection among the same layer. An RBM is an energy-based undirected generative model [13] that has a two-layer architecture in which the visible binary stochastic units $v \in \{0,1\}^D$ are connected to hidden binary stochastic units $h \in \{0,1\}^K$ using symmetrically weighted connections as shown in Fig. 1.

The energy of the state $\{v,h\}$ is defined as:

$$E(v,h; \theta) = -\sum_{i=1}^{D} \sum_{j=1}^{K} v_i W_{ij} h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{K} c_j h_j$$

where $v \in V$ is the observation nodes, $h \in H$ is the hidden random variables and $\theta = \{W,b,c\}$ are the model parameters: $W_{ij}$ is the symmetric interaction term between unit $i$ in the visible layer and unit $j$ in the hidden layer. $b_i$ is the $i$th bias of visible layer and $c_j$ is the $j$th bias of hidden layer. The probabilistic semantics for an RBM is defined by its energy function as follows:

$$p(v; \theta) = \frac{1}{Z(\theta)} \sum_{h \in H} \exp(-E(v,h; \theta))$$

$$Z(\theta) = \sum_{v \in V} \sum_{h \in H} \exp(-E(v,h; \theta))$$

where $Z(\theta)$ is the normalizing constant or partition function. The conditional distributions over hidden unit $h$ and visible vector $v$ are given:

$$p(h|v) = \prod_{j} p(h_j|v), \quad p(v|h) = \prod_{i} p(v_i|h)$$

the probability of turning on unit $j$ is a logistic function of the states of $v$ and $W_{ij}$:

$$p(h_j = 1|v) = \text{sigm} \left( \sum_{i} W_{ij} v_i + c_j \right)$$

the probability of turning on unit $i$ is a logistic function of the states of $h$ and $W_{ij}$:

$$p(v_i = 1|h) = \text{sigm} \left( \sum_{j} W_{ij} h_j + b_i \right)$$

where the logistic function is

$$\text{sigm}(x) = 1/(1 + \exp(-x))$$

The average of the log-likelihood with respect to the parameters $W$ can be retrieved from the Contrastive Divergence (CD) method [22]

$$\frac{\partial \ln p(v)}{\partial W_{ij}} = \langle v_i h_j \rangle_{P_{data}} - \langle v_i h_j \rangle_{P_{Model}}$$

where $\langle \cdot \rangle_{P_{data}}$ is an expectation with the data distribution and $\langle \cdot \rangle_{P_{Model}}$ is a distribution of samples from running the Gibbs sampler. In this case the term $\langle \cdot \rangle_{1}$ will be used such that it indicates an expectation with the distribution of samples from running the Gibbs sampler initialized at the data for one full step.
Then the parameter $W$ can be adjusted by:

$$W_{ij} = \theta W_{ij} + \eta (v_i h_j^t_{\text{data}} - v_i h_j^t)$$  \hspace{1cm} (9)$$

where $\theta$ is the momentum and $\eta$ is the learning rate.

The above description is based on one sample datum. In our research, we use all labeled data by inputting them one by one from the first layer. The deep architecture is constructed layer by layer from bottom to top and the weight is trained by calculated data in the $k$-1th layer.

2.3 Deep Belief Network

Deep Belief Network (DBN) is a generative model consisting of multiple stacked levels of neural networks that each layer contains a set of binary or real-valued units. The main building block networks for the DBN are restricted Boltzmann machine (RBM) layers and a backpropagation (BP) layer. An example of a DBN for classification is shown in Fig. 2. It consists of an input layer which contains the input units or called visible units, the hidden layers and finally an output layer that has one unit for each class. There is a full connections between two adjacent layers, but no two units in the same layer are connected.

To construct a DBN we train a stack of RBMs as many as the number of hidden layers in the DBN. First, the lowest layer is trained directly on the input data, so-called feature vector, then training each next higher layer by capturing the important feature of the hidden units of the previous layer as the input data for the next higher layer in order to get the weights in each RBM layer. RBM manages the feature vectors only excluding the label information. This procedure is continued until a number of hidden layers in the DBN have been trained unsupervisedly. Each RBM layer learns the parameters independently and makes the optimal parameters locally not for entire model. To address this optimization problem, there is a supervised BP layer on top of the model that fine-tunes the entire model in the learning process and generates the output in the inference process. Finally, the feature vector composes of some complicated features which reflect the structured information among the original features. When the stack of RBMs is trained, it can be used to initialize a multilayer neural network for classification tasks. The classification performance with the new feature vector is better than using the original feature vector.

3. Classification using DBN

The aim of this study is to use DBN in the classification of DNA microarray data like leukemia dataset. To train the DBN, we have used the training procedure as Algorithm 1.

Algorithm 1. Algorithm of deep belief network

**Input:**
- data $x,y$
- number of hidden units $N$
- number of layers $G$
- number of epochs $Q$
- number of labeled data $L$
- hidden layer $h$
- weight $W$
- biases of hidden layer $b$
- biases of visible layer $c$
- momentum $\theta$
- learning rate $\eta$


Output: model of deep architecture
for $k = 1; k < G$ do
for $r = 1; p \leq E$ do
for $s = 1; s \leq L$ do
\[ p(h_{j,s} = 1|v) = \text{sign} \left( \sum_i W_{ij}v_{i,s} + c_j \right) \]
\[ p(h_{j,s} = 1|v) = \text{sign} \left( \sum_i W_{ij}v_{i,s} + c_j \right) \]
update the weight and biases
\[ W_{ij} = \partial W_{ij} + \eta \left( (v_{i,s}h_{j,s})_{p_{\text{data}}} - (v_{i,s}h_{j,s})_{p_{t}} \right) \]
\[ b_i = b_i + \eta \left( (h_{j,s})_{p_{\text{data}}} - (h_{j,s})_{p_{t}} \right) \]
\[ c_j = c_j + \eta (v_t - v_t) \]
end
end
end

4. Experimental

4.1 Experimental setting

The proposed method has been evaluated by microarray datasets, which are lung cancer and prostate cancer in the first experiment before using leukemia data. The DBN successfully demonstrated the best test classification accuracies of 99.45% and 94.85% for lung cancer and prostate cancer dataset, respectively as shown in Table 1. Therefore, we adopted the DBN on the leukemia dataset. Leukemia data contains the expression levels of 7129 genes of 72 patients with two classes, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The training dataset consisted of 38 bone marrow samples with 27 ALL and 11 AML from adult patients. The testing dataset consisted of 24 bone marrow samples as well as 10 peripheral blood specimens from adults and children (20 ALL and 14 AML). Table 2 describe the specification of the leukemia dataset.

We compared the classification performance of DBN with the three representative classifiers, support vector machines (SVM), k-nearest neighbor (KNN) and naive bayes (NB) SVM and NB are the powerful classification methods and KNN is a typical nonlinear classifier that used as the standard for performance comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>DBN</th>
<th>SVM</th>
<th>KNN</th>
<th>NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer</td>
<td>99.45</td>
<td>98.90</td>
<td>94.46</td>
<td>97.80</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>94.85</td>
<td>91.18</td>
<td>81.62</td>
<td>55.15</td>
</tr>
</tbody>
</table>

4.2 Experimental results on leukemia data

The DBN classification has been applied on leukemia dataset. The performance of DBN was compared with SVM, KNN and NB by computing the accurate of the classification. We trained the weights of each layer with the number of epochs equal to 50 and the learning rate equal to 1.5 while the initial momentum is 0.5.

As shown in Table 3, the DBN classifier can obtain the classification accuracy of 98.61% for testing set of leukemia dataset. This shows that the DBN outperforms classifiers based SVM, KNN and NB algorithm.

5. Conclusion

DNA microarray experiment with a precise and reliable data analysis method can be used for cancer diagnosis. One of the methods that could help identifying the new pattern effectively is using machine learning approaches. In this paper we focused on the use of Deep Belief Network and applies it successfully to leukemia classification. We studied the details of DBN training and evaluated the performance of our approach on DNA microarray of leukemia data. Our result were compare with SVM, KNN and NB. The comparative results indicate that DBN has a better performance than the other classifiers for our data. The further work, we will consider gene selection such as Principle Component Analysis (PCA), Information Gain (IG) and Correlation-based Feature Selection (CFS) work with the DBN for improving the performance of leukemia classification and comparing with the other cancer datasets.

Acknowledgements

The authors would like to thank the Research and Development Office (RDO) of Prince of Songkla University (PSU), Thailand for their grant support. The first author would also like to express her special appreciations to the Graduates School of PSU for the graduate student research grant. Special thanks go to Mr.Pitipol Meemak of the Department of Molecular Biotechnology and Bioinformatics, Faculty of Science, PSU for his initiation in this research.
References


