Backpropagation with Diversive Curiosity: An Automatic Conversion From Search Stagnation to Exploration

Qun Dai*

Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China

Abstract. This paper proposes a novel approach, namely, the Back-propagation with Diversive Curiosity (DCPROP) algorithm, for solving the “flat spot” problem and for escaping from local minima. Representing the diversive curiosity, an internal indicator is designed for BP algorithm, which detects the phenomenon of being trapped in local minima and the occurrence of premature convergence. Upon such detection, the neural network is activated again to explore optimal solution in search space and escape from local minima by means of stochastic disturbance. The proposed DCPROP algorithm is implemented and applied to two well-known face recognition problems, and the results are compared with Standard Back-propagation (SBP).

Keywords: Diversive curiosity (DC); Standard Back-propagation (SBP); Back-propagation with Diversive Curiosity (DCPROP); Neural network; Face recognition.

1 Introduction

In our common opinions, curiosity is closely related to scientific discovery, exploration of the earth and the universe, interest in the supernatural, puzzles, spectator sports, and murder mysteries, etc[1]. We generally think of curiosity as a favorable trait motivating our human being’s learning activities. And it has also been consistently recognized as a critical motive which influences human behavior in both positive and negative ways at all stages of the life cycle. Famous ancient philosopher Aristotle and Cicero equated curiosity with a passion for learning in their works. So far, two well-known taxonomies about the types of curiosity are presented. The first one is given by William James in his classic work “The Principles of Psychology”, where he described two types of curiosity: “instinctual curiosity response” and “scientific curiosity”[1]. In instinctual curiosity response, attention is aroused by seeing something new, whereas in scientific curiosity, the “brain responds to an inconsistency or a gap in its knowledge, just as the musical brain responds to a discord in what it hears.” Scientific curiosity has been cited as a major impetus behind scientific discovery, possibly eclipsing even the drive for economic gain. The second taxonomy is given by Daniel Berlyne,
who divide curiosity into two types: “diversive curiosity” and “specific curiosity”. Diversive
curiosity (DC) mainly refers to the general tendency of a person to seek novelty, take risks, or
search for adventure, while specific curiosity describes the tendency to investigate a specific
object or problem for understand it thoroughly[1].

Complexity, novelty, uncertainty and conflict are the four major characteristics of external
stimuli which might arouse our curiosities[1]. In the pedagogical literature, curiosity has been
identified as a driving force in child development and as one of the most important spurs to
educational attainment. Teachers are encouraged to stimulate curiosity and are provided with
practical guidelines for doing so. Furthermore, in the commercial realm, curiosity is regarded
as a significant response evoked by literature and art, which can be exploited by advertisers to
harness the power of curiosity in elaborate designed advertisements[1].

In machine learning and optimization algorithm research community, Hong Zhang and
Masumi Ishikawa proposed a new method, Particle Swarm Optimization with Diversive
Curiosity (PSO/DC), which integrate the mechanism of diversive curiosity with PSO
algorithm for preventing premature convergence and for managing the exploration-exploitation trade-off[2]. Particle Swarm Optimization (PSO) is a stochastic and
population-based adaptive optimization algorithm, which has been widely applied to various
research fields in science and engineering including applications to large-scale, highly
nonlinear, and multimodal optimization problems[3, 4]. It was proposed by Kennedy and
Eberhart, inspired by the social behavior of flocks of birds and schools of fish[3, 4]. Its main
characteristics include particle dynamics, information exchange, intrinsic memory, and
directional search.

Although PSO model has good search performance with moderate computational cost and
accuracy, they tend to be trapped in local minima in solving multimodal optimization
problems. This phenomenon of premature convergence is the major obstacle that hinders the
improvement of the efficiency of PSO. To overcome this difficulty, Hong Zhang and Masumi
Ishikawa propose Particle Swarm Optimization with Diversive Curiosity (PSO/DC)[2]. Their
idea is to introduce a mechanism of diversive curiosity into PSO for preventing premature
convergence and for managing the exploration-exploitation trade-off. Diversive curiosity is
represented by an internal indicator that detects marginal improvement of a swarm of particles
for certain number of iterations, and forces them to continually explore an optimal solution to
a given optimization problem. The authors demonstrated the effectiveness of their approach
through applications to a 2-dimensional optimization problem. In their experiments, the
search performance of original EPSO was vastly improved by the mechanism of diversive
curiosity[2].

Back-propagation (BP) learning algorithm is a widely applied supervised learning
technique for training multi-layer feed-forward neural networks (MFN)[5]. BP algorithm has
many significant advantages, such as simplicity, ease of implementation and low computational complexity, etc. However, its convergence rate is slow, and it is easily trapped into local minima, especially for non-linearly separable and non-stationary problems\[6, 7\]. Most of these weaknesses are due to the “flat spot” problem. The so-called “flat spot” problem mainly refers to the occurrence of premature saturation in the derivative of the activation function. It results in slow learning speed, and suppressed weight update process of neural model. Many modifications have been proposed to improve the performance of BP, many of which focus on solving the “flat spot” problem to increase the convergence speed\[8-16\]. The Resilient Back-Propagation (RPROP) learning algorithm is proposed to perform a local adaptation of the weight-updates according to the behavior of the error function to overcome the inherent disadvantages of pure gradient-descent\[9\]. The Simulated Annealing Resilient Back-Propagation (SARPROP) algorithm is an enhancement to RPROP based on simulated annealing\[10\]. It mainly involves the combination of weight constraints with noise early in training to force the network to perform a more thorough search of the initial weight space, and then, the network is allowed to refine its solutions when training procedure continues. The Levenberg-Marquardt algorithm (LMA)\[11\] is an integration of gradient descent technique and the Gauss-Newton algorithm (GNA)\[17\] to speed up the convergence process. The Back-propagation with Magnified Gradient Function (MGFPROP)\[12\] algorithm is proposed to increase the convergence rate and the global convergence capability by magnifying the gradient function of the activation function without violating the BP algorithm gradient descent property. Although these improved BP algorithms increase the convergence rate of the original BP, they still suffer from the “flat spot” problem and still sometimes trap in local minima.

We propose a novel approach called Back-propagation with Diversive Curiosity (DCPROP) in this paper, by integrating the mechanism of diversive curiosity with BP learning algorithm for solving the “flat spot” problem and for escaping from local minima. As explained beforehand, the mechanism of diversive curiosity here refers to a concept in psychology, i.e., the general tendency of seeking novelty, taking risks, or searching for adventure. In their work, Hong Zhang and Masumi Ishikawa proposed Particle Swarm Optimization with Diversive Curiosity (PSO/DC)\[2\], by introducing a mechanism of diversive curiosity into PSO. In their algorithm, an internal indicator is crucial, which detects premature convergence of a swarm of particles, and provides information to make them active to explore the global solution in search space. They implemented the mechanism of diversive curiosity by introducing this critical internal indicator. As a result, PSO/DC can successfully prevent premature convergence, and manage the exploration-exploitation trade-off. They verified the effectiveness of their PSO/DC through applications to a 2-dimensional multimodal optimization problem\[2\]. The inspiration of this paper originates from Hong Zhang and
Masumi Ishikawa’s research of PSO/DC algorithm. Imitating their implementation approach of the mechanism of diversive curiosity through an internal indicator in PSO/DC algorithm, an internal indicator is designed for the Standard Back-propagation (SBP) learning algorithm, which detects the phenomenon of being trapped in local minima and the occurrence of premature convergence. Upon detection of such phenomenon, the MFN network is able to escape from local minima with the use of stochastic disturbance, and become active again to explore better solution in its search space, thus the “flat spot” problem is partially solved and the convergence rate of SBP algorithm is increased to some extent. The proposed novel modified BP algorithm in this work is entitled Back-propagation with Diversive Curiosity, i.e. DCPROP, for short.

In DCPROP, apart from the above mentioned design of an internal indicator and the integration of diversive curiosity mechanism with the standard BP algorithm, two small novel techniques are put forth for the final implementation of DCPROP. The first one is the design of a random disturbance term (RDT) based on Gaussian white noise, i.e. proper RDT is added to the network weights upon detection of premature convergence in the training process. The amplitude of RDT is designed to be linearly decreased with the increase of training iteration, thus with the training process entering into endgame, the RDTs will become small enough to allow the final convergence of the whole algorithm. The second small technique is, in order to enhance the learning capability of DCPROP and, at the same time, preserve the final convergence property of the original BP algorithm, with every 100 times of the stochastic disturbances operations to the network weight vectors, the standard BP training iterations are subsized by one.

Therefore, as a whole, the main contributions of this paper are: first, the mechanism of diversive curiosity is introduced into SBP for its performance optimization; second, an internal indicator is designed for the implementation of diversive curiosity in SBP, which detects the phenomenon of being trapped in local minima and the occurrence of premature convergence; third, a proper random disturbance term (RDT) is designed and added to the network weights upon the detection of premature convergence, whose amplitude is designed to be linearly decreased with the increase of training iteration to allow the final convergence of the whole algorithm; forth, with each 100 stochastic disturbances operations to the network weights, the scheduled total training iterations of SBP are subsized by one, so that the learning capability of DCPROP could be improved and, at one time, the final convergence property of SBP algorithm is still kept down.

In our computer experiments, we implement DCPROP algorithm to two famous facial recognition problems, with the first group of experiments performed on the Olivetti Research Laboratory (ORL) face database\cite{18}, and the second one on the Yale face database\cite{19}. Experimental results comparisons are provided among the proposed DCPROP algorithm, SBP
and other state-of-the-art algorithms.

This paper is organized as follows. Section 2 briefly introduces the standard BP algorithm. Section 3 describes the curiosity concept, the proposed internal indicator for simulation the mechanism of diversive curiosity, and the proposed Backpropagation with Diversive Curiosity (DCPROP) algorithm in detail. Section 4 presents applications of DCPROP to two facial recognition problems, describes the implementation results and provides comparisons to SBP and other state-of-the-art approaches. Finally, Section 5 gives conclusions and suggests further works.

2 The Standard BP (SBP) algorithm

The basic structure of the considered three-layered (Feed-forward Neural Network) FNN is as shown in Fig.1. The network possesses $N$ input nodes, $H$ hidden nodes, and $M$ output nodes, and adopts sigmoidal function as its activation function.

![The basic structure of a three-layered FNN](image)

The procedure of the standard BP (SBP) algorithm is as follows:

01: Begin

02: Initialize network weights vectors $\mathbf{\tilde{w}}$ and $\mathbf{w}$. Set the learning rate $\eta$, the momentum factor $\alpha$, and the error threshold $\epsilon_1$ to small positive values, respectively. Set the maximum iteration number $I$. Set $i = 0$, $d_1 = -1$.

03: While $i \leq I$ and $d_1 < 0$ Do

04: For each training sample $x_p = (x_{p1}, \cdots, x_{pn})$, $p = 1, \cdots, P$ Do

05: Input a training sample $x_p = (x_{p1}, \cdots, x_{pn})$ to the network, and compute
the output vector $\tilde{o}_p$ and $o_p$ layer by layer using the following equations:

$$
\tilde{o}_{ph} (i) = f\left( \sum_{n=1}^{N} \tilde{w}_{nh} (i) x_{pn} \right) \quad (1)
$$

$$
o_{pm} (i) = f\left( \sum_{h=1}^{H} w_{hm} (i) \tilde{o}_{ph} (i) \right) \quad (2)
$$

where $f(\cdot)$ is adopted as the sigmoid function:

$$
f(x) = \frac{1}{1 + e^{-x}} \quad (3)
$$

06: End For

07: Calculate the sum squared error $E(i)$ for the $i$-th training iteration as follows:

$$
E(i) = \frac{1}{2} \sum_{p=1}^{P} \sum_{m=1}^{M} [t_{pm} - o_{pm} (i)]^2 \quad (4)
$$

where $t_p = (t_{p1}, \cdots, t_{pm})$ is the target vector corresponding to the $p$-th training sample $x_p$.

08: If $E(i) < \varepsilon_1$ Then $d_i = 1$.

09: Else calculate the network weights changes for the next iteration $\Delta \tilde{w}_{nh} (i+1)$ and $\Delta w_{hm} (i+1)$ as follows:

$$
\Delta w_{hm} (i+1) = -\eta \frac{\partial E(i)}{\partial w_{hm} (i)} + \alpha \Delta w_{hm} (i)
$$

$$
= \eta \sum_{p=1}^{P} \delta_{pm} (i) \tilde{o}_{ph} (i) + \alpha \Delta w_{hm} (i) \quad (5)
$$

$$
\Delta \tilde{w}_{nh} (i+1) = -\eta \frac{\partial \tilde{o}_p (i)}{\partial \tilde{w}_{nh} (i)} + \alpha \Delta \tilde{w}_{nh} (i)
$$

$$
= \eta \sum_{p=1}^{P} \tilde{d}_{ph} (i) x_{pn} + \alpha \Delta \tilde{w}_{nh} (i) \quad (6)
$$

where

$$
\delta_{pm} (i) = (t_{pm} - o_{pm} (i)) o_{pm} (i)(1 - o_{pm} (i)) \quad (7)
$$
\begin{align*}
\tilde{\delta}_p(i) &= \tilde{\delta}_p(i)(1 - \tilde{\delta}_p(i)) \sum_{m=1}^{M} \delta_{pn}(i) w_{mn}(i) \\
\end{align*}

10: \quad \text{Update the network weights vectors for the } (i+1)\text{-th iteration as follows:} \quad w(i+1) = w(i) + \Delta w(i+1) \quad \text{(9)}

and \quad \tilde{w}(i+1) = \tilde{w}(i) + \Delta \tilde{w}(i+1) \quad \text{(10)}

11: \quad i = i + 1 ;
12: \quad \text{End If;}
13: \quad \text{End While;}
14: \quad \text{End}

3 \quad \textbf{Backpropagation with Diversive Curiosity (DCPROP) Algorithm}

3.1 \quad \textbf{Diversive Curiosity}

Curiosity is tightly related to scientific attainments, discovery and imagination. It is also relevant to technological innovation, contribution and revolution. It leads to exploration of ourselves, the earth and the universe. It is the inner source of our interest in the supernatural, puzzles, sports, and mysteries, etc. Aristotle and Cicero equated curiosity with a passion for learning\textsuperscript{[1, 20]}. And many great men in history also proposed their own viewpoints about curiosity:

Curiosity is insubordination in its purest form.—from Bend Sinister, by Vladimir Nabokov\textsuperscript{[1]}.

Curiosity is, in great and generous minds, the first passion and the last— Samuel Johnson\textsuperscript{[1]}.

Loewenstein noted that “curiosity occupies a critical position at the crossroads of cognition and motivation” and that “educators know much more about educating motivated students than they do about motivating them in the first place\textsuperscript{[20].}.”

Burns and Gentry state that “identification of manageable knowledge gaps that complement the natural curiosity in a learner, combined with explicit connections to the learner’s value system, will generate tension-to-learn in the learner\textsuperscript{[21].}.”

Just out of curiosity, do we have a Plan B?—Nathan Lane as Preed in Titan A.E. \textsuperscript{[1]}

Curiosity is the thing. If you don’t give life to curiosity, you haven’t done your job.—Carlos Picon\textsuperscript{[1]}.

In William James’ classic work “The Principles of Psychology”, which was published in 1890, he described two types of curiosity. The first one is an instinctual or emotional response, where concern is inspired by seeing something new. The second one is so-called “scientific
“curiosity” or “metaphysical wonder” where the “brain responds to an inconsistency or a gap in its knowledge, just as the musical brain responds to a discord in what it hears”.

However, in the early twentieth century, Daniel Berlyne in his research proposed another classification of curiosity, i.e. diversive and specific curiosity. Diversive curiosity (DC) refers to as a general tendency for a person to seek novelty, take risks, or search for adventure, while specific curiosity (SC) is referred to a tendency to investigate a specific problem in detail so as to understand it more thoroughly. It was further determined by Berlyne that curiosity can be aroused by external stimuli with characteristics such as complexity, novelty, uncertainty and conflict. The intensity of stimulation should be moderate, neither too high nor too low. Too high-level stimulation will result in anxiety, while too low-level stimulation will cause lack of motivation. Only when the level of stimulation is just right, it will result in exploratory behavior. This phenomenon was described by Day as the “Zone of Curiosity”.

3.2 Internal Indicator for BP Algorithm

Hong Zhang and Masumi Ishikawa proposed a novel Particle Swarm Optimization with Diversive Curiosity (PSO/DC) algorithm in their paper, where they incorporated diversive curiosity mechanism into the original PSO. Specifically speaking, in their algorithm they designed a special internal indicator, which can detect the occurrence of premature convergence of a swarm of particles. Upon the detection of premature convergence, they made the swarm of particles execute re-initialization, so as to force them to explore better solution in search space. They concluded that their PSO/DC can successfully prevent premature convergence and manage the exploration-exploitation trade-off. The effectiveness of their PSO/DC was verified through application to a 2-dimensional multimodal optimization problem. Enlightened by Hong Zhang and Masumi Ishikawa’s proposal, an internal indicator is designed for the standard Back-propagation (SBP) learning algorithm in our paper imitating their designment of PSO/DC. Our internal indicator can decide the occurrence of being trapped in local minima and the happening of premature convergence of BP algorithm. When such phenomenon is discovered by the internal indicator, the neural system is managed to escape form local minima by means of stochastic disturbance, making it recover active, so that it will explore better solution in its search space. Consequently, the so-called “flat spot” problem is partially solved and the convergence rate of SBP algorithm is increased to some degree.

We term the improved BP algorithm here as Backpropagation with Diversive Curiosity (DCPROP) algorithm. Loewenstein ever stated that “diversive curiosity occupies a critical position at the crossroad of cognition and motivation”. Hong Zhang and Masumi Ishikawa interpreted “cognition” as the act of precisely locating a solution (exploitation), and
“motivation” as the intention of exploring the global solution (exploration). According to them, exploitation is due to execution of specific curiosity, and exploration is due to execution of diversive curiosity. Namely, a swarm of particles will carry out a conversion from specific curiosity to diversive curiosity at the position for the dissatisfaction to present situation such as premature convergence and search stagnation. We find their ideas reasonable and valuable. Similarly, we propose the following internal indicator, $\beta$, for determining premature convergence of SBP algorithm and helping it escape flat spots.

$$\beta_i(J, \varepsilon_2) = \max(\varepsilon_2 - \sum_{j=1}^{J} \frac{|E(i) - E(i-j)|}{J}, 0)$$

where $E(i)$ refers to the error function of SBP as defined in Eq.(4), $J$ refers to the duration of judgment for SBP, and $\varepsilon_2$ is also a positive tolerance parameter for detection of premature convergence, similar to PSO/DC.

From Eq. (11), it can be seen that, if the value of the internal indicator, $\beta$, equals zero, it means the error function value for the $i$-th training iteration is still remarkably changing, however, when $\beta$ turns into positive value, it means the error function value for the $i$-th training iteration is no longer continuously changing greatly. In the proposed DCPROP algorithm, if $\beta$ becomes positive, appropriate stochastic disturbances are added to network weights. In this way, we solve the “flat spot” problem of the original BP algorithm to some degree and increase its convergence rate accordingly. Under the effects of stochastic disturbances, the neural network is forced to become active again, and starts to explore better solution in its search space. As result, it escapes form a certain amount of local minima. This result is attributed to the function of diversive curiosity mechanism. Because, similar to PSO/DC, it is the usage of internal indicator that achieves automatic conversion of the DCPROP algorithm from search stagnation to exploration stage during gradient descent search. Therefore, it is concluded that the design idea of $\beta_i$ implements the mechanism of diversive curiosity within the proposed novel DCPROP algorithm.

3.3 The Random Disturbance Term (RDT) in Design

White noise is a random signal (or process) with a flat power spectral density. In other words, the signal contains equal power within a fixed bandwidth at any center frequency. White noise draws its name from white light in which the power spectral density of the light...
is distributed over the visible band in such a way that the eye's three color receptors (cones) are approximately equally stimulated\textsuperscript{[25,26]}.

The Random Disturbance Term (RDT) based on Gaussian white noise is designed as follows:

\[ RDT(i) = wgn(M, N, P) \cdot (C_1 / (i + C_2)) \]  \hspace{1cm} (12)

where \( RDT(i) \) refers to the random disturbance term of the \( i \)-th training iteration. Function \( wgn(M, N, P) \) generates an \( M \)-by-\( N \) matrix of white Gaussian noise, and \( P \) specifies the power of the output noise in dBW. \( C_1 \) and \( C_2 \) are two adjustable parameters, which are assigned to 5 and 500 respectively in our experiments. The design idea of RDT lies in: proper RDT is added to the network weights when needed, i.e. when premature convergence is detected during training process. However, the amplitude of RDT is required to be linearly decreased with the increase of training iterations, because with the training process entering into endgame, the RDTs are demanded to be small enough to allow the final convergence of the whole algorithm.

3.4 Backpropagation with Diversive Curiosity (DCPROP) algorithm

The network internal indicator \( \beta_i \) estimates the emergence of premature convergence, and then by means of appending random disturbance terms to the network weights, the neural system is activated again, and is able to flee from those “flat spots”, continuing to explore other solutions in its subsequent search process.

The procedure of Backpropagation with Diversive Curiosity (DCPROP) algorithm is shown below:

01: Begin
02: Initialize network weights vectors \( \hat{\mathbf{w}} \) and \( \mathbf{w} \). Set the learning rate \( \eta \), the momentum factor \( \alpha \), the error threshold \( \varepsilon_1 \) and tolerance parameter \( \varepsilon_2 \) for network premature convergence to small positive values, respectively. Set the maximum iteration number \( I \). Set current iteration number \( i = 0 \). Set the counter \( RDT \_ Count = 0 \), which counts the frequency of adding Random Disturbance Term (RDT). Set boolean variables \( d_1 = \text{false} \) and \( d_2 = \text{false} \).
While \( i \leq I \) and \( d_i \) is false Do

If \( d_2 \) is true Then

Add appropriate RDTs to network weights \( \hat{\mathbf{w}} \) and \( \mathbf{w} \);

\[ RDT\_Count = RDT\_Count + 1; \]

If \( \text{mod}(RDT\_Count, 100) = 0 \)

\[ I = I + 1; \]

End If

Else

For each training sample \( \mathbf{x}_p = (x_{p1}, \ldots, x_{pN}) \), \( p = 1, \ldots, P \) Do

Input a training sample \( \mathbf{x}_p = (x_{p1}, \ldots, x_{pN}) \) to the network, and compute the output vector \( \hat{\mathbf{o}}_p \) and \( \mathbf{o}_p \) layer by layer using the following equations:

\[
\hat{o}_{ph}(i) = f\left(\sum_{n=1}^{N} \hat{w}_{nh}(i)x_{pn}\right)
\]

\[
o_{pm}(i) = f\left(\sum_{h=1}^{H} w_{hm}(i)\hat{o}_{ph}(i)\right)
\]

where \( f(\cdot) \) is adopted as the sigmoid function:

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

End For

Calculate the sum squared error \( E(i) \) for the \( i \)-th training iteration as follows:

\[
E(i) = \frac{1}{2} \sum_{p=1}^{P} \sum_{m=1}^{M} \left( t_{pm} - o_{pm}(i) \right)^2
\]

where \( \mathbf{t}_p = (t_{pl1}, \ldots, t_{plM}) \) is the target vector corresponding to the \( p \)-th training sample \( \mathbf{x}_p \).

If \( E(i) < \epsilon_i \) Then \( d_i = 1; \)

Else calculate the network weights changes for the next iteration.
\( \Delta \tilde{w}_{mh}(i+1) \) and \( \Delta w_{hm}(i+1) \) as follows:

\[
\Delta w_{hm}(i+1) = -\eta \frac{\partial E(i)}{\partial w_{hm}(i)} + \alpha \Delta w_{hm}(i)
\]

\[
= \eta \sum_{p=1}^{P} \delta_{pm}(i)\tilde{o}_{ph}(i) + \alpha \Delta w_{hm}(i)
\]

\[
\Delta \tilde{w}_{mh}(i+1) = -\eta \frac{\partial E(i)}{\partial \tilde{w}_{mh}(i)} + \alpha \Delta \tilde{w}_{mh}(i)
\]

\[
= \eta \sum_{p=1}^{P} \tilde{\delta}_{ph}(i)x_{pm} + \alpha \Delta \tilde{w}_{mh}(i)
\]

where

\[
\delta_{pm}(i) = (t_{pm} - o_{pm}(i))o_{pm}(i)(1 - o_{pm}(i))
\]

\[
\tilde{\delta}_{ph}(i) = \tilde{o}_{ph}(i)(1 - \tilde{o}_{ph}(i))\sum_{m=1}^{M} \delta_{pm}(i)w_{hm}(i)
\]

12: Update the network weights vectors for the \((i+1)\)-th iteration as follows:

\[
w(i+1) = w(i) + \Delta w(i+1)
\]

and

\[
\tilde{w}(i+1) = \tilde{w}(i) + \Delta \tilde{w}(i+1)
\]

13: End If

14: End If

15: Compute the value of network internal indicator \( \beta_i \);

16: If \( \beta_i \leq 0 \) Then \( d_z = false \);

17: Else \( d_z = true \);

18: End If

19: \( i = i + 1 \);

20: End While;

21: End

3.5 An Important Detailed Issue

With respect to the design idea of this algorithm, there is one more important detail issue needed to be clarified. In the algorithm, a counter \( RDT\_Count \) is utilized to count the
times that the program enters into step 04, which means the frequency that the if statement “$d_2$ is true” is satisfied, and also means the frequency that appropriate RDTs are appended to network weights instead of standard weights adjustments of the original BP algorithm. With the increase of the total $RDT \_Count$, the standard BP training iterations might be too less to realize its sufficient learning. Therefore, with 100 times of the increase of the total $RDT \_Count$, we subsidize the standard BP training iterations by one. As result, firstly, the occurrence of being trapped in local minima and premature convergence is timely detected. Secondly, the neural network is activated again, escaping from local minima and continuing to explore better solution in search space. At last, the learning capability and final convergence property of the original BP algorithm are still kept for the proposed DCPROP algorithm.

3.6 Discussions about the Time Complexity of DCPROP algorithm

According to the detailed procedure of DCPROP algorithm listed out in Section 3.4, although with every 100 times of the stochastic disturbances operations to the network weight vectors, the standard BP training iterations are subsidized by one, however, both the decrease of Random Disturbance Terms (RDT) and the convergence of the algorithm are much faster than the increase of the total training iterations, so that the final convergence property of the whole algorithm is not affected and still holds. Therefore, as a result, the running time of DCPROP algorithm is still $O(I)$, which is of the same order of SBP in time complexity, and the calculation rapidity of DCPROP algorithm is also very close to that of SBP.

4 Experimental Results and Comparisons

The task of face recognition plays an important role in human life. The way we get along with other people is firmly based on our abilities to properly recognize them. Face recognition is an effortless task for human being. With a quick glance at a face, we are able to recognize the face and, most of the time, name the person. Such a process takes place so easily and instantaneously that we ourselves do not know exactly what happens in our mind when we try to recognize one facial image. One of the most striking characteristics of human facial identification is its robustness\cite{27}. Human beings are able to identify distorted images\cite{28}, coarse images, partially occluded faces, and even inverted face images\cite{29}.

4.1 Experiments on ORL Face Database

The first group of computer experiments are carried out for investigating the pattern
recognition performance of DCPROP for the facial recognition task based on the ORL face database[18] as shown in Fig. 2. It possesses 40 persons and 10 different images for each person, 92x112 pixels with slightly varying lighting conditions, pose, scale, face expression and presence or absence of glasses[30].

![Fig.2 Examples of ORL face image database](image)

In experiments one hidden layer with 30 units has been used for the FNN. All the images have been scaled down by factor 2 to speed up learning. Since it is known from D. Bryliuk and V. Starovoitov’s research work that such scaling do not greatly influence recognition rate on the ORL database[18, 30]. The number of input units of FNN is set to be the number of image pixels, i.e. 46x56. Gray level of each pixel has been linearly scaled from range [0; 255] to [0; 0.1] for avoiding paralysis or surfeit of the network. The number of output units is set to be equal to the number of classes, i.e. 40, the number of persons in the ORL database. Each output unit corresponds to one unique class among these 40 classes.

![Training Set](image)

![Testing Set](image)

(5:5) Training Set                          Testing Set

(4:6) Training Set                          Testing Set

(3:7) Training Set                          Testing Set
In this group of experiments, three subgroups of experiments have been carried out: for the first subgroup, five images among the total ten images of each person are used for training, and the remaining five are used for testing; for the second subgroup, four images of each person are used for training, and the remaining six for testing; and finally, for the third subgroup, three images of each person are used for training, and the remaining seven for testing. Evidently, the less the ratio of training patterns number to testing ones is, the more challenging the task will be\textsuperscript{[31]}. Fig. 3 illustrates the three different training/testing set partition strategies for one of the 40 persons in ORL database.

Table 1 FNN Trained with SBP or DCPROP Algorithm Final Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>FNN Trained with SBP</th>
<th>FNN Trained with DCPROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Units Number</td>
<td>2576 (i.e. 46*56)</td>
<td>2576</td>
</tr>
<tr>
<td>Hidden Units Number</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Output Units Number</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.135</td>
<td>0.295</td>
</tr>
<tr>
<td>Momentum Item</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Minimum Error</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>Initial Weights Values Range</td>
<td>-1 to 1</td>
<td>-1 to 1</td>
</tr>
<tr>
<td>Maximum Iterations</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>Judgment Duration $J$</td>
<td>-</td>
<td>170</td>
</tr>
<tr>
<td>Tolerance Parameter $\epsilon_2$</td>
<td>-</td>
<td>$3.33 \times e^{-4}$</td>
</tr>
</tbody>
</table>

Table 2 Recognition Error Rates with Three Different Proportions of Partition for Training and Testing Set for 30 repetitive runnings, respectively.

<table>
<thead>
<tr>
<th>Partition Proportions for Training and Testing Set</th>
<th>Recognition Error Rates for FNN Trained with SBP</th>
<th>Recognition Error Rates for FNN Trained with DCPROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>5:5</td>
<td>10%</td>
<td>8.5%</td>
</tr>
<tr>
<td>4:6</td>
<td>16%</td>
<td>11.07%</td>
</tr>
<tr>
<td>3:7</td>
<td>19%</td>
<td>14.17%</td>
</tr>
</tbody>
</table>

Experiments have been carried out intended to test the performance of the proposed DCPROP learning algorithm and compare its performance with the standard BP learning algorithm. Training and testing both algorithms was implemented with the following system configuration: 2.6-GHz PC with 1 GB of RAM using Windows XP operating system,
MATLAB-language source code.

Table 1 lists the final parameters of the successfully trained FNNs utilizing DCPROP and SBP learning algorithms, respectively. The first seven parameters (except for Learning Rate) are equal for both algorithms to ensure a relatively objective comparison. Table 2 shows the corresponding recognition error rates using different training set and testing set partitioning ratios. From the results shown in Table 2, it can be easily observed that DCPROP outperforms SBP conformably for all the three kinds of partitioning methods for training and testing set. The use of 200 images for training and the remaining 200 images for testing the generalized neural network has been selected in the remaining experiments of this work, because such ratio provides the least recognition error rates.

![Fig.4 The performance indices of DCPROP with two adjustable parameters, i.e. judgment](image)
interval $J$ and tolerance parameter $\varepsilon_2$. (a) The facial recognition error rates, (b) the number of $\textit{RDT}_\textit{Count}$.

For estimating appropriate values for judgment duration $J$ and tolerance parameter $\varepsilon_2$ in the internal indicator $\beta_i$, and investigating the faces recognition performance of DCPROP algorithm, we change the values of parameters in the indicator, i.e., judgment duration $J = 80, 90, 100, 110, 120, 130, 140, 150, 160, 170$ and tolerance parameter $\varepsilon_2 = 10^{-3}, 5 \times 10^{-4}, 3.33 \times 10^{-4}, 2.5 \times 10^{-4}$ (i.e. $\varepsilon_2 = 0.01/10, 0.01/20, 0.01/30, 0.01/40$).

Actually, at the early stage of the experiments, we select tolerance parameter $\varepsilon_2 = 0.01/10, 0.01/10^2, 0.01/10^3, 0.01/10^4, 0.01/10^5$. As result, the numbers of $\textit{RDT}_\textit{Count}$ turn out to be zeros for the latter three values, which indicate that the corresponding designated $\varepsilon_2$ values are too small. Thus, we come up with the above relatively suitable $\varepsilon_2$ values, which are adjusted a little larger.

Fig.4 demonstrates the experimental results, i.e., the facial recognition error rates, and the number of $\textit{RDT}_\textit{Count}$ of DCPROP for each case of the two adjustable parameters, i.e. judgment interval $J$ and tolerance parameter $\varepsilon_2$. From the experimental results, the following features of DCPROP can be observed.

1. According to the facial recognition error rates shown in Fig.4 (a), the recommended values for the judgment duration $J$ are 110, 120, 140 and 170.
2. From the results shown in Fig.4 (b), with the increment of the judgment duration $J$, the number of $\textit{RDT}_\textit{Count}$ decrease nonlinearly in general. And it is easy to observe that, the choice of the four $\varepsilon_2$ values is decisive for the number of $\textit{RDT}_\textit{Count}$. The less the $\varepsilon_2$ value is, the less the number of $\textit{RDT}_\textit{Count}$ will be, and vice versa. This phenomenon might be interpreted as the sensitivity of our DCPROP system. The larger the $\varepsilon_2$ value is, the more sensitive the DCPROP system becomes to the occurrence of premature convergence and “flat spots” problem.
3. Also, according to the facial recognition error rates shown in Fig.4 (a), the recommended
values of the tolerance parameter $\varepsilon_2$ are $5 \times 10^{-4}$ and $2.5 \times 10^{-4}$. Actually, the proper combination of judgment duration $J$ and tolerance parameter $\varepsilon_2$ is crucial. As shown in Fig.4 (a), the four pairs of combination of $J$ and $\varepsilon_2$, i.e., $J = 80$ and $\varepsilon_2 = 3.33 \times 10^{-4}$, $J = 140$ and $\varepsilon_2 = 3.33 \times 10^{-4}$, $J = 150$ and $\varepsilon_2 = 3.33 \times 10^{-4}$ and $J = 170$ and $\varepsilon_2 = 3.33 \times 10^{-4}$ obtain the four least facial recognition error rates. However, improper combination of $J$ and $\varepsilon_2$, such as $J = 130$ and $\varepsilon_2 = 3.33 \times 10^{-4}$, results in the second largest error rates.

Fig.5 The comparison of facial recognition error rates between SBP and DCPROP algorithm for 30 repetitive runnings.

Fig.5 shows the comparative results of faces recognition error rates between SBP and DCPROP algorithm for 30 trials, adopting the most suitable parameters for both, as listed out in Table 1. Compared to the error rates obtained by SBP algorithm, those obtained by the proposed DCPROP algorithm are smaller in scale, significantly more stationary and stable. The above experimental results confirm the effectiveness of the internal indicator in implementing the mechanism of diversive curiosity and integrating the diversive curiosity mechanism with SBP algorithm, although the experimental data are not very good for those improper combinations of $J$ and $\varepsilon_2$. 
Other experimental results performed on the ORL face database in recent 10 years are listed out in Table 3. Although compared to the recognition error rates shown in Table 3, our DCPROP algorithm is not in the ascendant at all. However, the main intention of this work is to propose an interesting idea of introducing the notion of diverse curiosity into SBP algorithm, so as to lessen its problems of local minima and premature convergence. Similar ideas could be applied to other algorithms to optimize their performances further, which might be the aims of our future work. It might not be very equitable and significative to compare the recognition performances of DCPROP and other state-of-the-art methods directly and absolutely. Therefore, the facts that DCPROP algorithm receives better performances in comparison with SBP could be a desirable result for this work.

4.2 Experiments on Yale Face Database

Yale face database comprises 165 GIF images of 15 persons. There are 11 images per person, one for each of the following facial expressions or configurations: center-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised, and wink.

In this group of experiments, similar to but somewhat different from the above one, three subgroups of experiments have been performed: for the first subgroup, six images among all the eleven images of each person are utilized for training, and the remaining five are utilized for testing; for the second subgroup, seven images of each person are used for training, and the remaining four for testing; and finally, for the third subgroup, eight images of each person are used for training, and the remaining three for testing. Naturally, the less the ratio of training patterns number to testing ones is, the more challenging the task will be. Table 4 lists out the average recognition error rates and standard deviations of SBP and DCPROP algorithm for 30 trials, for the three subgroups of experiments respectively, adopting the most suitable parameters for both, as listed out in Table 1. In comparison to the experimental results of SBP, those achieved by DCPROP algorithm are obviously smaller in error rates, and significantly more stable according to the standard deviations obtained.

Other results recently performed on the Yale database, i.e. average recognition error rates

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recognition Error Rate (%)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDBNN[33]</td>
<td>4</td>
<td>1997</td>
</tr>
<tr>
<td>Point-Matching[34]</td>
<td>16</td>
<td>1998</td>
</tr>
<tr>
<td>PCA+RBF[35]</td>
<td>4.9</td>
<td>2000</td>
</tr>
<tr>
<td>Wavelet+RBF[36]</td>
<td>3.7</td>
<td>2001</td>
</tr>
</tbody>
</table>
and standard deviations obtained by other state-of-the-art algorithms are shown in Table 5, respectively. It can be observed easily that DCPROP achieved better performances in this database compared to other algorithms. Consequently, the effectiveness of DCPROP algorithm is verified by this group of experiments.

### Table 4: Average Recognition Error Rates and Standard Deviation (%) with Three Different Proportions of Partition for Training and Testing Set for 30 repetitive runnings, respectively.

<table>
<thead>
<tr>
<th>Partition Proportions for Training and Testing Set</th>
<th>SBP</th>
<th>DCPROP</th>
</tr>
</thead>
<tbody>
<tr>
<td>6:5</td>
<td>24.53±2.36</td>
<td>21.16±1.67</td>
</tr>
<tr>
<td>7:4</td>
<td>21.94±5.76</td>
<td>15.28±2.7</td>
</tr>
<tr>
<td>8:3</td>
<td>16.59±3.18</td>
<td>15.19±1.76</td>
</tr>
</tbody>
</table>

Table 5: Other results recently performed on the Yale database (Average Recognition Error Rates and Standard Deviation (%))[^37]

<table>
<thead>
<tr>
<th></th>
<th>SpCCA</th>
<th>PSemi-RS</th>
<th>HSemi-RS</th>
<th>PRS_SpCCA</th>
<th>HRS_SpCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6</td>
<td>32.73±4.46</td>
<td>25.49±3.77</td>
<td>28.0±2.51</td>
<td>23.27±2.91</td>
<td>25.82±2.34</td>
</tr>
<tr>
<td>G5</td>
<td>38.33±0.57</td>
<td>29.0±2.7</td>
<td>30.7±2.4</td>
<td>27.33±0.59</td>
<td>30.31±2.21</td>
</tr>
</tbody>
</table>

5 Conclusion

This paper introduces a new concept in neural networks with the aim of motivating more theoretical and practical developments on the integration of diversive curiosity in cognitive modeling. The idea of simulating diversive curiosity and utilizing it to improve the learning and the decision making of standard BP algorithm is investigated. SBP is found to be limited in performance due to the local minima and “flat spot” problems. To overcome these problems, this paper proposes a new approach called Backpropagation with Diversive Curiosity (DCPROP) algorithm. The key idea of the approach is an internal indicator designed for the standard BP algorithm, which detects the phenomenon of being trapped in local minima and the occurrence of premature convergence. Upon such detection, the MFN network is activated over again to explore optimal solution in search space and escape form local minima by means of stochastic disturbance. The mechanism of this approach can be interpreted as the mechanism of diversive curiosity. With regard to the time complexity of DCPROP algorithm, its running time is the same with SBP, i.e. $O(I)$.

The main contributions of the paper are as follows:

1. The mechanism of diversive curiosity is introduced into SBP for its performance
optimization.

(2) An internal indicator is designed for the implementation of diversive curiosity in SBP, which detects the phenomenon of being trapped in local minima and the occurrence of premature convergence.

(3) A proper random disturbance term (RDT) is designed, whose amplitude linearly decreases with the increase of training iterations, thus permitting the final convergence of the whole algorithm.

(4) For each 100 times of the stochastic disturbances operations, the SBP training iterations is subsidized by one to simultaneously improve the learning capability and keep the final convergence property for DCPROP.

Applications of DCPROP to two famous faces recognition problem well demonstrated its effectiveness. So far, the activation to the MFN network is only attempted by means of stochastic disturbance. In future work, other activation methods will be tried. Also so far, only two face recognition problems was carried out for demonstrating the effectiveness of DCPROP. It is left for further study to apply DCPROP to other high-dimensional benchmark problems, and to other complex application problems in the real-world. And finally, similar ideas of optimizing algorithm performances with the help of the diversive curiosity mechanism could be applied to some other algorithms, which might be the aims of our future work as well.

Acknowledgements

We thank the editors and reviewers of the paper for their valuable comments. This work is supported by the National Natural Science Foundation of China under the Grant no. 61100108, and Ministry of Education NUAA research project under the Grant no. NS2011015, and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References


