A Continuous Restricted Boltzmann Machine with a Hardware-Amenable Learning Algorithm

Hsin Chen and Alan Murray

Dept. of Electronics and Electrical Engineering,
University of Edinburgh, Mayfield Rd.,
Edinburgh, EH9 3JL, UK,
{Hsin.Chen,A.F.Murray}@ee.ed.ac.uk

Abstract. This paper proposes a continuous stochastic generative model that offers an improved ability to model analogue data, with a simple and reliable learning algorithm. The architecture forms a continuous restricted Boltzmann Machine, with a novel learning algorithm. The capabilities of the model are demonstrated with both artificial and real data.

1 Introduction

Probabilistic generative models offer a flexible route to improved data modelling, wherein the stochasticity represents the natural variability of real data. Our primary interest is in processing and modelling analogue data close to a sensor interface. It is therefore important that such models are amenable to analogue or mixed-mode VLSI implementation.

The Product of Experts (PoE) has been shown to be a flexible architecture and "Minimising Contrastive Divergence" (MCD) can underpin a simple learning rule [1]. The Restricted Boltzmann Machine (RBM) [2] with an MCD rule has been shown to be amenable to further simplification and use in real applications [3]. The RBM has one hidden and one visible layer with only inter-layer connections. Let s_i and s_j represent the states of the stochastic units i, j, and w_{ij} be the interconnect weights. The MCD rule for RBM replaces the computationally-expensive relaxation search of the Boltzmann Machine with:

$$\Delta w_{ij} = \eta (< s_i s_j >_0 - < \hat{s_i} \hat{s_j} >_1) \quad (1)$$

$\hat{s_i}$ and $\hat{s_j}$ correspond to one-step Gibbs sampled "reconstruction" states, and $<>$ denotes expectation value over the training data. By approximating the probabilities of visible units as analogue-valued states, the RBM can model analogue data [1][3]. However, the binary nature of the hidden unit causes the RBM to tend to reconstruct symmetric analogue data only, as will be shown in Sect. 3.

The rate-coded RBM (RBMrate) [4] removes this limitation by sampling each stochastic unit for m times. The RBMrate unit thus has discrete-valued states, while retaining the simple learning algorithm of (1). RBMrate offers an improved ability to model analogue image [4], but the repetitive sampling will cause more spiking noise in the power supplies of a VLSI implementation, placing the circuits in danger of synchronisation [5].

J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415, pp. 358–363, 2002.

2 The Continuous Restricted Boltzmann Machine

2.1 A Continuous Stochastic Unit

Adding a zero-mean Gaussian with variance σ^2 to the input of a sampled sigmoidal unit produces a continuous stochastic unit as follows:

$$s_j = \varphi_j \left(\sum_i w_{ij} s_i + \sigma \cdot N_j(0,1) \right), \tag{2}$$

$$\text{with} \quad \varphi_j(x_j) = \theta_L + (\theta_H - \theta_L) \cdot \frac{1}{1 + \exp(-a_j x_j)} \tag{3}$$

where $N_j(0,1)$ represents a unit Gaussian, and $\varphi_j(x)$ is a sigmoid function with lower and upper asymptotes at θ_L and θ_H, respectively. Parameter a_j controls the steepness of the sigmoid function, and thus the nature of the unit's stochastic behaviour. A small value of a_j renders input noise negligible and leads to a near-deterministic unit, while a large value of a_j leads to a binary stochastic unit. If the value of a_j renders the sigmoid linear over the range of the added noise, the probability of s_j remains Gaussian with mean $\sum_i w_{ij} s_i$ and variance σ^2. Replacing the binary stochastic unit in RBM by this continuous form of stochastic unit leads to a continuous RBM (CRBM).

2.2 CRBM and Diffusion Network

The model and learning algorithms of the Diffusion Network (DN) [6][7] arise from its continuous stochastic behaviour, as described by a stochastic differential equation. A DN consists of n fully-connected units and an $n \times n$ real-valued matrix W, defining the connection-weights. Let $x_j(t)$ be the state of neuron j in a DN. The dynamical diffusion process is described by the Langevin equation:

$$dx_j(t) = \kappa_j \left(\sum_i w_{ij} \varphi_i(x_i(t)) - \rho_j x_j(t) \right) \cdot dt + \sigma \cdot dB_j(t) \tag{4}$$

where $1/\kappa_j > 0$ and $1/\rho_j > 0$ represent the input capacitance and resistance of neuron j. $dB_j(t)$ is the Brownian motion differential [7]. The increment, $B_j(t + dt) - B_j(t)$, is thus a zero-mean Gaussian random variable with variance dt. The discrete-time diffusion process for a finite time increment Δt is:

$$x_j(t + \Delta t) = x_j(t) + \kappa_j \sum_i w_{ij} \varphi_i(x_i(t)) \Delta t - \kappa_j \rho_j x_j(t) \Delta t + \sigma z_j(t) \sqrt{\Delta t} \tag{5}$$

where $z_j(t)$ is a Gaussian random variable with zero mean and unit variance. If $\kappa_j \rho_j \Delta t = 1$, the terms in $x_j(t)$ cancel and writing $\sigma\sqrt{\Delta t} = \sigma'$, this becomes:

$$x_j(t + \Delta t) = \kappa_j \sum_i w_{ij} \varphi_i(x_i(t)) \Delta t + \sigma' z_j(t) \tag{6}$$

If $w_{ij} = w_{ji}$ and κ_j is constant over the network, the RHS of (6) is equivalent to the total input of a CRBM as given by (2). As $s_j = \varphi_j(x_j)$, the CRBM is simply a symmetric restricted DN (RDN), and the learning algorithm of the DN is thus a useful candidate for the CRBM.

2.3 M.C.D. Learning Algorithms for the CRBM

The learning rule for the parameter λ_j of the DN is [6]:

$$\Delta\lambda_j =< S_{\lambda_j} >_0 - < S_{\lambda_j} >_\infty \tag{7}$$

where $<>_0$ refers to the expectation value over the training data with visible states clamped, and $<>_\infty$ to that in free-running equilibrium. S_{λ_j} is the system-covariate [6], the negative derivative of the DN's energy function w.r.t. parameter λ_j. The restricted DN can be shown to be a PoE [8] and we choose to simplify (7) by once again minimising contrastive divergence [1].

$$\Delta\hat{\lambda_j} =< S_{\lambda_j} >_0 - < \hat{S}_{\lambda_j} >_1 \tag{8}$$

where $<>_1$ indicates the expectation values over one-step sampled data. Let $\varphi(s)$ represent $\varphi_j(s)$ with $a_j = 1$. The energy function of CRBM can be shown to be similar to that of the continuous Hopfield model [9][6].

$$U = -\frac{1}{2}\sum_{i \neq j} w_{ij} s_i s_j + \sum_i \frac{\rho_i}{a_i}\int_0^{s_i} \varphi^{-1}(s)ds \tag{9}$$

(8) and (9) then lead to the MCD learning rule for the CRBM's parameters:

$$\Delta\hat{w_{ij}} = \eta_w(< s_i s_j >_0 - < \hat{s_i}\hat{s_j} >_1) \tag{10}$$

$$\Delta\hat{a_j} = \eta_a \left(\frac{\rho_j}{a_j^2} \left\langle \int_{\hat{s_j}}^{s_j} \varphi^{-1}(s)ds \right\rangle \right) \tag{11}$$

where $\hat{s_j}$ denotes the one-step sampled state of unit j, and $<>$ in (11) refers to the expextation value over the training data. To simplify the hardware design, we approximate the integral term in (11) as

$$\int_{\hat{s_j}}^{s_j} \varphi^{-1}(s)ds \propto (s_j + \hat{s_j})(s_j - \hat{s_j}) \tag{12}$$

The training rules for w_{ij} and a_j thus require only adding and multiplying calculation of local units' states.

3 Demonstration: Artificial Data

Two-dimensional data were generated to probe and to compare the performance of RBM and CRBM on analogue data (Fig.1(a)). The data include two clusters

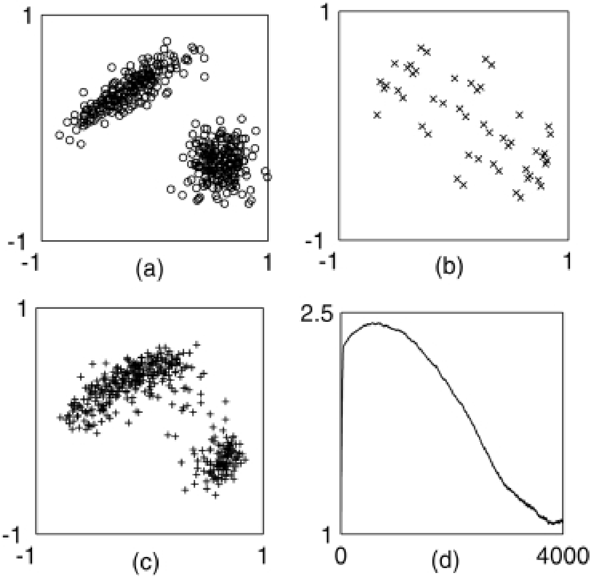

Fig. 1. (a) Artificial-generated analogue training data. (b) Reconstruction by the trained RBM (c) Reconstruction by the trained CRBM (d)Learning trace of a_j.

of 200 data. Figure 1(b) shows points reconstructed from 400 random input data after 20 steps of Gibbs' sampling by an RBM with 6 hidden units, after 4000 training epochs. The RBM's tendancy to generate data in symmetric patterns is clear. Figure 1(c) shows the same result for a CRBM with four hidden units, $\eta_w = 1.5$, $\eta_a = 1$ and $\sigma = 0.2$ for all units. The evolution of the gain factor a_j of one visible unit is shown in Fig.1(d) and displays a form of 'autonomous annealing', driven by (11), indicating that the approximation in (12) leads to sensible training behaviour in this stylised, but non-trivial example.

4 Demonstration: Real Heart-Beat (ECG) Data

To highlight the improved modelling richness of the CRBM and to give these results credence, a CRBM with four hidden units was trained to model the ECG data used in [3] and [10]. The ECG trace was divided into one training dataset of 500 heartbeats and one test dataset of 1700 heartbeats, each of 65 samples. The 500 training data contain six Ventricular Ectopic Beats (VEB), while the 1700 test data contain 27 VEBs. The CRBM was trained for 4000 epochs with $\eta_w = 1.5$, $\eta_a = 1$, $\sigma = 0.2$ for visible units and $\sigma = 0.5$ for hidden units.

Figure 2 shows the reconstruction by the trained CRBM, from initial visible states as 2(a) an observed normal QRS complex 2(b) an observed typical VEB, after 20 subsequent steps of unclamped Gibbs' sampling. The CRBM models both forms of heartbeat successfully, although VEBs represent only 1% of the

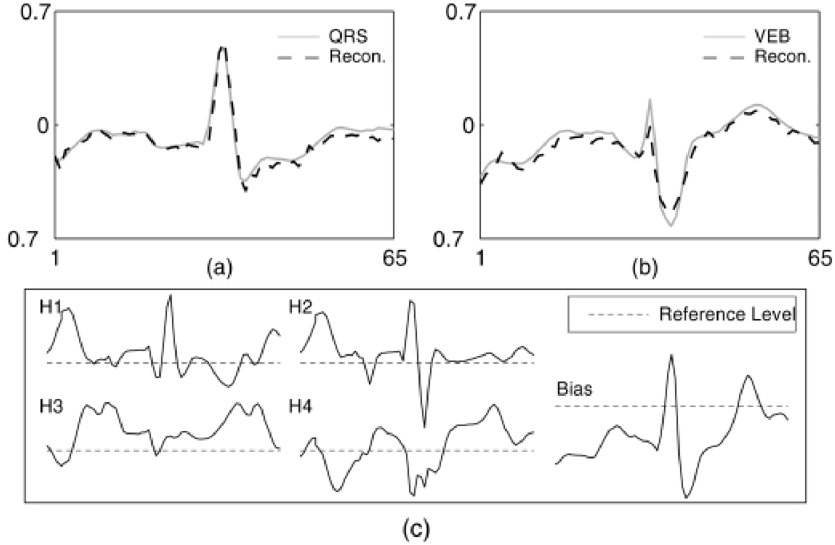

Fig. 2. Reconstruction by the trained CRBM with input of (a) a normal QRS and (b) a typical VEB. (c) The receptive fields of the hidden bias and the four hidden units

training data. Following [3], Fig.2(c) shows the receptive fields of the hidden bias unit and the four hidden units. The Bias unit codes an "average" normal QRS complex, and H3 adds to the P- and T- waves. H1 and H2 drive a small horizontal shift and a magnitude variation of the QRS complex. Finally, H4 encodes the significant dip found in a VEB. The most principled detector of VEBs in test data is the log-likelihood under the trained CRBM. However, log-likelihood requires complicated hardware. Figure 2(c) suggests that the activities of hidden units may be usable as the basis of a simple novelty detector. For example, the activities of H4 corresponding to 1700 test data are shown in Fig.3, with the noise source in equation (2) removed. The peaks indicate the VEBs clearly. VL in Fig.3 indicates the minimum H4 activity for a VEB and QH marks the datum with maximum H4 activity for a normal heartbeat. Therefore,

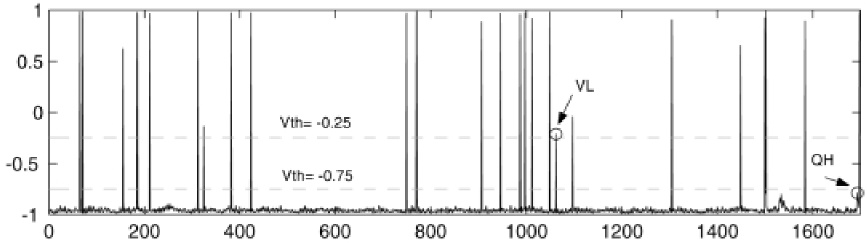

Fig. 3. The activities of H4 corresponding to 1700 test data.

even a simple linear classifier with threshold set between the two dashed line will detect the VEBs with an accuracy of 100%. The margin for threshold is more than 0.5, equivalent to 25% of the total value range. A single hidden unit activity in a CRBM is, therefore, potentially a reliable novelty detector and it is expected that layering a supervised classifier on the CRBM, to "fuse" the hidden unit activities, will lead to improved results.

5 Conclusion

The CRBM can model analogue data successfully with a simplified MCD rule. Experiments with real ECG data further show that the activities of the CRBM's hidden units may function as a simple but reliable novelty detector. Component circuits of the RBM with the MCD rule have been successfully implemented [5][11]. Therefore, the CRBM is a potential continuous stochastic model for VLSI implementation and embedded intelligent systems.

References

1. Hinton, G.E.: Training Products of Experts by minimising contrastive divergence. Technical Report: Gatsby Comp. Neuroscience Unit, no.TR2000-004. (2000)
2. Smolensky, P.: Information processing in dynamical systems: Foundations of harmony theory. Parallel Distributed Processing. Vol.1 (1986) 195–281
3. Murray, A.F.: Novelty detection using products of simple experts-A potential architecture for embedded systems. Neural Networks 14(9) (2001) 1257–1264
4. Teh, Y.W., Hinton, G.E.: Rate-coded Restricted Boltzmann Machine for face recognition. Advances in Neural Information Processing Systems, Vol.13 (2001)
5. Woodburn, R.J., Astaras, A.A., Dalzell, R.W., Murray, A.F., McNeill, D.K.: Computing with uncertainty in probabilistic neural networks on silicon. Proc. 2nd Int. ICSC Symp. on Neural Computation. (1999) 470–476
6. Movellan, J.R.: A learning theorem for networks at detailed stochastic equilibrium. Neural Computation, Vol.10(5) (1998) 1157–1178
7. Movellan, J.R., Mineiro, P., William, R.J.: A Monte-Carlo EM approach for partially observable diffusion process: Theory and applications to neural network. Neural Computation (In Press)
8. Marks, T.K., Movellan, J.R.: Diffusion networks, products of experts, and factor analysis. UCSD MPLab Technical Report, 2001.02. (2001)
9. Hopfield, J.J.: Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Nat. Academy of Science of the USA, Vol.81(10). (1984) 3088–3092
10. Tarassenko, L., Clifford G.: Detection of ectopic beats in the electrocardiogram using an Auto-associative neural network. Neural Processing Letters, Vol.14(1) (2001) 15–25
11. Fleury, P., Woodburn, R.J., Murray, A.F.: Matching analogue hardware with applications using the Products of Experts algorithm. Proc. IEEE European Symp. on Artificial Neural Networks. (2001)63–67