ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

B-bleaching: Agile Overtraining Avoidance in the WiSARD Weightless Neural Classifier

Danilo S. Carvalho[1], Hugo C. C. Carneiro[1], Felipe M. G. França[1], Priscila M. V. Lima[2] ▯

1- Universidade Federal do Rio de Janeiro - PESC/COPPE
Rio de Janeiro, Brazil

2- Universidade Federal Rural do Rio de Janeiro - PPGMMC/DEMAT
Seropédica, Brazil

Abstract. Weightless neural networks constitute a still not fully explored Machine Learning paradigm, even if its first model, WiSARD, is considered. Bleaching, an improvement on WiSARD's learning mechanism was recently proposed in order to avoid overtraining. Although presenting very good results in different application domains, the original sequential bleaching and its confidence modulation mechanisms still offer room for improvement. This paper presents a new variation of the bleaching mechanism and compares the three strategies performance on a complex domain, that of multilingual grammatical categorization. Experiments considered both number of iterations and accuracy. Results show that binary bleaching allows for a considerable improvement to number of iterations whilst not introducing loss of accuracy.

1 Introduction

As the areas of application of Artificial Intelligence expand, so do the demand for speed and accuracy in classification and training techniques. Moreover, many situations require that training be interleaved with classification, in an online learning fashion. Though not a recently proposed paradigm, weightless neural networks (WNNs) are still not fully explored [1]. WNNs first model, WiSARD [2], possesses the ability of performing online training. However, it often suffers from overtraining after a not so big set of examples. WiSARD's learning mechanism has been recently improved by the addition of a process called *bleaching* [3]. The original sequential bleaching and its confidence modulation mechanisms presented promising results in different application domains [3] [4]. There is still, however, room for improvement.

The following is how the remainder of the text is organised. Background knowledge on WiSARD and bleaching is briefly reviewed in Section 2. Section 3 presents a new variation of the bleaching mechanism and Section 4 provides a quantitative comparison of the three strategies performance. The problem of multilingual grammatical categorization of ambiguous words was the chosen complex domain chosen for the comparison. Section 5 provides some concluding remarks as well as points to future research steps.

▯This work was partially supported by CAPES, CNPq and FAPERJ Brazilian research agencies.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

2 WiSARD and Bleaching

WiSARD (Wilkie, Stonham & Aleksander's Recognition Device)[2] is a weightless neural network formed by various RAM-discriminators, each consisting of a set of X one-bit word RAMs with n address inputs. This way, the network receives a binary pattern of $X \times n$ bits as input. All RAM address lines are connected to the input pattern by means of a biunivocal pseudo-random mapping, and all of the RAM contents are initially set to zero.

The training is done by setting to "1" the memory locations addressed by the input patterns. WiSARD classifies unseen patterns by summing the memory contents addressed thereof and thus obtaining the number of RAMs that output "1". Such number is called r, the *discriminator response*, which expresses a similarity measure of the input pattern with respect to the patterns in the training set. Each RAM-discriminator is associated with a particular class, so that when a pattern is given as input, each RAM-discriminator gives a response r to that input. The various RAM-discriminator responses are evaluated by an algorithm which compares them and computes the relative confidence c of the highest response, as illustrated in Figure 1.

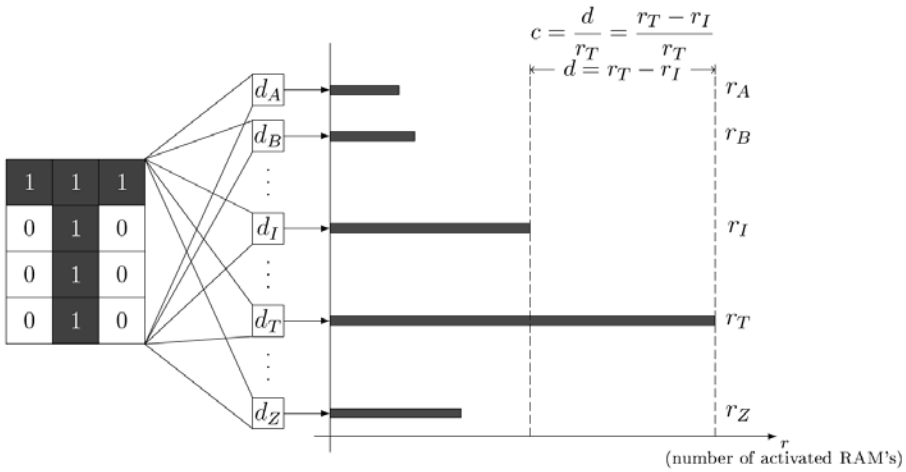

Fig. 1: Example of multidiscriminator responses r and confidence calculus.

Similar to other conventional or weightless neural network systems, RAM-based networks are not immune to overtraining, specially on noisy training data. As different patterns are presented to the network, the memory locations addressed by these patterns are set to "1". If the training set has many patterns, most of the RAM positions tend to be addressed and set to "1", meaning that RAM-discriminators will output high values of r for any given input pattern, thus increasing the probability of obtaining ties during the classification phase. This effect is called *saturation* of the RAM-neurons, and undermines the generalisation capabilities of the network. Saturation can be solved by noticing

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

that representative patterns should occur more often than others in the training data. Therefore, the addressing frequency of the RAM locations should reveal which parts of the stored pattern (i.e., sub-patterns) are relevant for calculating similarity with respect to the training set. This observation was absorbed by DRASiW [5] [6], an extension to WiSARD which employs access counters in order to register the amount of times a location is addressed, instead of the original one bit memory locations. What is left is to find a way of isolating the relevant sub-patterns from the others. This role is filled by a technique called *bleaching* [3]. Bleaching uses the frequency information from DRASiW to eliminate RAM-discriminator response ties in the following way: (i) a *bleaching threshold* variable $b \geq 1$ is set; (ii) all memory locations with access count greater than or equal to b are set to "1"; remaining ones are set to "0"; Figure 2 presents a snapshot of a RAM-discriminator response having $b = 2$.

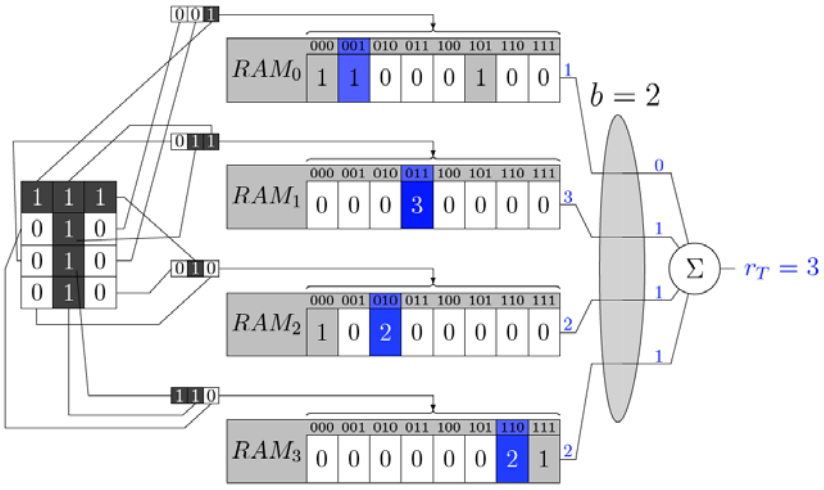

Fig. 2: Example of RAM-discriminator d_T; $X = 4$, $n = 3$, $b = 2$.

Starting with $b = 0$, the threshold is increased while there are ties in the discriminators' responses. When there's no more ties, the class chosen is the one whose discriminator has the highest output. The effect of bleaching on overtraining consists of making the RAM-neurons ignore the sub-patterns that the network considered too uncommon, i.e., the ones that were presented to it less than b times. This procedure leaves only the relevant sub-patterns and thus solves the saturation problem efficiently. However, bleaching has to be carried out observing that if b is too high, only the most frequent sub-patterns are retained, and generalisation is lost; if it is too low, saturation can still persist. Besides that, eliminating the ties on WiSARD does not guarantee that the chosen output was either the correct or an optimal one (i.e., high confidence score).

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

3 Bleaching styles

The performance of *bleaching* can be optimized either by minimizing the amount of time to eliminate ties or by maximizing the relative confidence c of the chosen response. The first issue can be seen as a search problem. [7] proposes a binary search approach to find the optimum b. Alternatively, the second issue requires knowledge about how the discriminators' responses change as b increases. It depends on the network and input characteristics, such as the number of RAM-neurons, inputs, training examples and the input data nature.

In this paper we investigate the behavior of the discriminator response as the *bleaching threshold* is increased and compare three bleaching algorithms, aiming to cope with these issues. The method used in this study can be broken into two steps, the graphical analysis and the algorithm evaluation.

During the first step, the network is trained and then several samples are presented for it to classificate. For each sample classified, the saturated RAM-discriminators are submitted to a simple form of *bleaching*, in which b is increased until all the discriminators output zero. Each b increment yields a set of pairs (r_i, b), where r_i is the response of the ith discriminator. Afterwards, a graphical representation is built, by using each pair as a point. The result is a set of "response curves" that decrease as b increases. This representation is used to understand the behavior of discriminator response and to plan algorithms that leverage response characteristics. Figure 3 presents an example of the graphical representation.

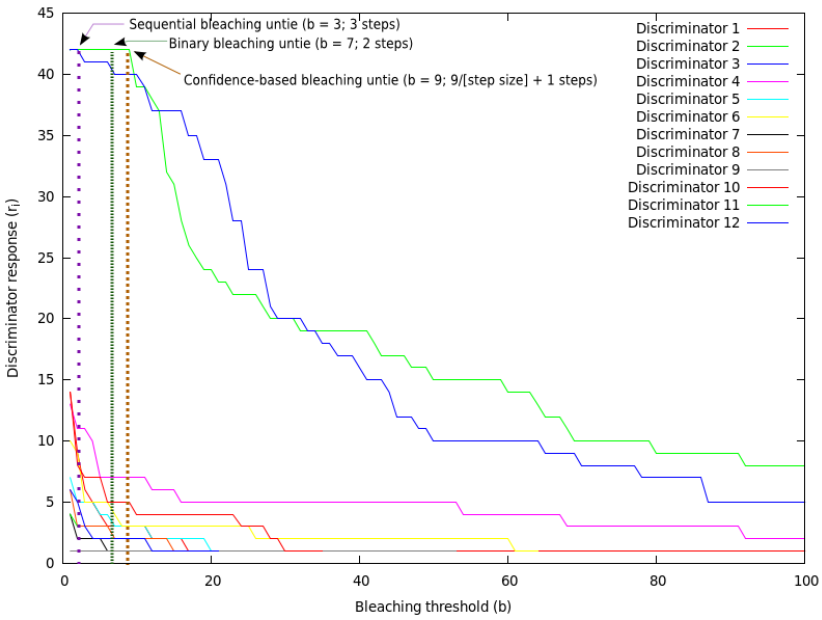

Fig. 3: Typical discriminator response behavior.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

In the following step, algorithms are used to solving the two issues above mentioned. They are then evaluated taking into account its performance measurements (e.g. time, accuracy and response confidence) and then improved by matching its solution with the graphical representation. The algorithms evaluated this way were: *a)* Sequential search: increases b by a unit until the ties are eliminated with a response confidence greater than a threshold d; *b)* Confidence search [4]: similar to the sequential search, but with variable increment for b. It stops at the first maximum confidence value, and *c)* Binary search: does a binary search on b, $b \in [1, b_{max}]$, in which b_{max} is the highest value in some memory position of a discriminator. It uses the geometric mean in place of the arithmetic mean. The search is stopped when it finds a b value for which there are no ties and the highest response value is the same as with $b = 1$.

The data used for the experiments was obtained from the mWANN-Tagger [8], a WiSARD-based multilingual part-of-speech tagger. 2000 samples were pseudo-randomly chosen for each type. Each dataset used in the experiments was split into 10 subsets. Afterwards, a 10-fold cross-validation procedure was performed and 200 samples were pseudo-randomly chosen in each fold. Only ambiguous words were selected to be part of these samples. This way, more ties would occur, thus causing the algorithms to be more intensively exercised. For each sample classified, the saturated state of the network was written into a file as an entry containing an identifier for the correct class and a list of the values in the discriminators' RAMs, one discriminator per line. The network configurations and the languages chosen are shown in Table 1.

Language	Number of RAMs	Inputs per RAM
Chinese	84	37
English	27	36
Portuguese	42	38
Turkish	14	19

Table 1: Network configurations.

The evaluation was based in the following metrics: time, accuracy and response confidence. Time was measured by counting the number of iterations, i.e., how many times b was changed until the solution was found; accuracy by asserting whether the solution was the correct response or not, and the response confidence by storing the one obtained for each solution. The results of the evaluation process are summarized in Table 2. This table present the mean value obtained for each of the measures. Their standard deviation values were also calculated, but their values were extremely small (For the time it was usually 10 times smaller than the mean, and 100 times for the time).

4 Conclusion

A novel variation of the bleaching mechanism was presented in this paper: b-bleaching. Its performance was compared to that of the two previous versions of

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Language	Measure	Sequential	Confidence-based	Binary
Chinese	Accuracy	0.878	0.893	0.892
	Time	16.562	71.646	**1.217**
English	Accuracy	0.921	0.924	0.927
	Time	4.264	47.075	**1.267**
Portuguese	Accuracy	0.959	0.962	0.961
	Time	7.327	218.273	**1.258**
Turkish	Accuracy	0.818	0.825	0.821
	Time	24.191	148.608	**2.299**

Table 2: Performance of bleaching threshold searching algorithms.

bleaching with respect to both number of iterations and accuracy. Experiments on the hard and complex domain of multilingual grammatical categorization of ambiguous words suggest that binary bleaching hastens the classification step by an order of 3.365 to 13.609 compared to the sequential algorithm, and 37.155 to 173.508 compared to the confidence-based one. Further work include the application of the three versions of bleaching to other domains and the use of an observer network to perform adjustments of the algorithms parameters.

References

[1] I. Aleksander, M. De Gregorio, F. M. G. França, P. M. V. Lima, and H. Morton. A brief introduction to weightless neural systems. In *Proc. of ESANN 2009*, page 6–12. i6doc.com, April 2009.

[2] I. Aleksander, W. V. Thomas, and P. A. Bowden. WISARD: A radical step foward in image recognition. *Sensor Review*, 4:120–124, July 1984.

[3] B. P. A. Grieco, P. M. V. Lima, M. De Gregorio, and F. M. G. França. Producing pattern examples from "mental" images. *Neurocomputing*, 73(7–9):1057–1064, March 2010.

[4] C. R. Souza, F. F. Nobre, P. M. V. Lima, R. M. Silva, R. M. Brindeiro, and F. M. G. França. Recognition of HIV-1 subtypes and antiretroviral drug resistance using weightless neural networks. In *Proc. of ESANN 2012*, page 429–434. i6doc.com, April 2012.

[5] Massimo de Gregorio. On the reversibility of multi-discriminator systems. Technical Report 125/97, Istituto di Cibernetica–CNR, 1997.

[6] C. M. Soares, C. L. F. da Silva, M. De Gregorio, and F. M. G. França. Uma implementação em software do classificador WISARD (In Portuguese). In *V Simpósio Brasileiro de Redes Neurais*, volume 2, page 225–229, Belo Horizonte, MG, Brazil, December 1998.

[7] H. C. C. Carneiro, F. M. G. França, and P. M. V. Lima. WANN-Tagger: A weightless artificial neural network tagger for the Portuguese language. In *Proc. of ICFC-ICNC 2010*, page 330–335. SciTePress, October 2010.

[8] Hugo C. C. Carneiro. The role of the index of synthesis of languages in part-of speech tagging with weightless artificial neural networks (In Portuguese). Master's thesis, Universidade Federal do Rio de Janeiro, 2012.