# Neural network pruning using discriminative information for emotion recognition

M. Sánchez-Gutiérrez[1] and E. M. Albornoz[2]

[1] Departamento de Matemáticas Aplicadas y Sistemas
*Universidad Autónoma Metropolitana*, México
edmax86@gmail.com
[2] Instituto **sinc**(*i*)
*Universidad Nacional del Litoral - CONICET*, Argentina
emalbornoz@sinc.unl.edu.ar



**Abstract.** In the last years, the effort devoted by the scientific community to develop better emotion recognition systems has been increased, mainly impulsed by the potential applications. The Boltzmann restricted machines (RBM) and the deep machines of Boltzmann (DBM) are models that, in recent years, have received much attention due to their good performance for different issues. However, it is usually difficult to measure their predictive capacity and, specifically, the individual importance of hidden units. In this work, some measures are computed in the hidden units in order to rank their discriminative ability among multiple classes. Then, this information is used to prune those units that seem less relevant. The results show a significant decrease in the number of units used in the classification at the same time that the error rate is improved.

**Keywords:** RBM · DBM · pruning · entropy · divergence · feature selection · emotions.

## 1 Introduction

While humans can differentiate most of the natural emotions expressed in almost all environmental conditions, machine learning systems still present difficulties in this task. In recent years, a series of systems for automatic emotion recognition have been developed with varying degrees of success [21]. In the case of artificial neural networks, there are several criteria used to evaluate a network's quality e.g. training time, scalability, and generalisation ability, among others. One common approach to determine an appropriate network size for a specific task is by using heuristics and/or trial-and-error, usually looking for good performance and generalisation ability on a validation set. Another approach considers ways of 'growing' an artificial neural network until satisfactory performance is achieved [7,22]. A different technique uses 'pruning' methods [4,24,14]. In general, these methods begin by training an artificial neural network, which is large enough to ensure a satisfactory performance. Afterwards, neurons are removed from the trained net (for example, the ones with the smallest weights) and then the network is often fine-tuned or retrained. This procedure could also be repeated until

some convergence criterion is achieved, otherwise the smallest network that performed adequately is assumed to have the most suitable topology for the given data set. This type of pruning was called post–training pruning (PTP) [4].

Networks size is especially relevant and recent works show that larger or deeper nets can solve the tasks using a more appropriate space [15,8,13]. In consequence, new complications associated with complex and computationally demanding training algorithms must be addressed [23,12,3]. In this context, optimise a feed–forward artificial neural network has proven to be a difficult task. The best results obtained on supervised learning tasks involve an unsupervised learning component, usually in an unsupervised greedy pre–training phase [6,11].

In this work, the standard DBM-RBM configuration is considered, where a RBM is training (unsupervised) at the first step and then, a posterior classifier is feeding with its outputs. However, instead of using the last layer of the RBM to feed the classifier, the more discriminative hidden units are used based on a post-training ranking. In order to measure the discriminative capability two criterion were used and the multi-class emotion classification task was addressed. A binary approach was presented in [18].

In the next section, the material and methods are introduced. Section 3 deals with experiments and results and finally, the discussions are presented.

## 2 Materials and methods

As it was mentioned, the multi-class emotion classification task is addressed using two emotional speech corpora and well-known parameterisations.

### 2.1 Speech corpora and feature extraction

Both databases have been extendly used and they are labeled using seven emotions with a distribution showed in Table 1. From the INTERFACE project which involves four languages: English, French, Slovenian and Spanish, the last one was used here. This corpus was created by the Center for Language and Speech Technologies and Applications (TALP) of the Polytechnic University of Catalonia (UPC) with the purpose of investigating emotional discourse. Two professional actors, a man and a woman, elicited 5113 spoken sentences. The other corpus was developed at the Communication Science Institute, in the Berlin Technical University [2]. The corpus has 535 utterances, and the same sentences were recorded in German by 10 actors: 5 females and 5 males. In a first step, 10 utterances for each emotion type (from 1 to 7 sec.) were elicited and then, a perception test with 20 individuals was carried out to ensure the emotional quality and naturalness of the utterances and the most confusing were eliminated.

Although there is no a definitive consensus about the best characteristics for emotional speech recognition [5], the research community considers some suitable attributes to define baselines[20,25]. In this work, a well-known set computed over the whole sentences was used: the means of the {first 12 MFCCs, $F_0$, energy} and the zero-crossing rate; in addition, the means of their first derivatives. Consequently, each audio file is represented by a 30–dimensional vector.

Table 1: Emotional corpora distribution.

| | neutral | anger | disgust | fear | joy | surprise boredom$^{(*)}$ | sadness |
|---|---|---|---|---|---|---|---|
| INTERFACE | 734 | 724 | 731 | 735 | 731 | 728 | 730 |
| EmoDB | 79 | 127 | 46 | 69 | 71 | $81^{(*)}$ | 62 |

### 2.2 Classifiers: restricted Boltzmann machines

The RBM is an artificial neural network with two layers (Fig. 1): the input (visible) layer and the hidden (output) layer [8,6]. There is no connections between the units in the same layer [10], and the RBM represents the joint distribution between the input vector and hidden layers (random variables).

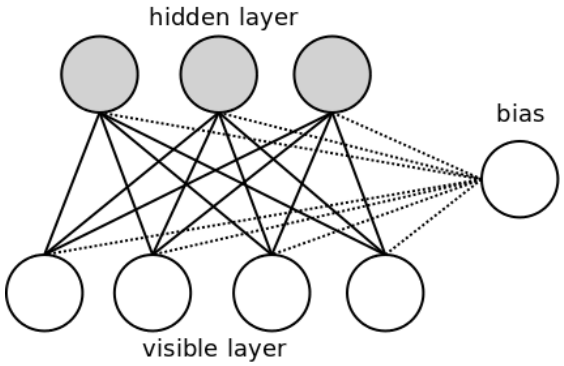



Fig. 1: Restricted Boltzmann machine

As it is a generative stochastic network, it can learn the probability distribution over the data using an energy function $E$ defined as:

$$E(v,h) = -a^\top v - b^\top h - v^\top W h \tag{1}$$

where $v$ and $h$ are the input and the hidden state vectors respectively, $W$ is a symmetric matrix of the connection weights, and $\{a, b\}$ are bias vectors for the layers. The joint probability ($p(v,h)$) assigns a probability to each configuration $(v,h)$ using:

$$p(v,h) = \frac{e^{-E(v,h)}}{Z} \tag{2}$$

where $Z = \sum_{v,h} e^{-E(v,h)}$ is a normalisation constant. Then, the probability assigned by the network to the visible vector $v$ is:

$$p(v) = \frac{1}{Z} \sum_{h} e^{-E(v,h)} \tag{3}$$

As there is no connections in the same layer, the visible variables are conditionally independent, given the hidden variables, and vice versa. Then, the conditional probabilities are: $p(v_j = 1|h) = \sigma(a_i + \sum_j h_j w_{i,j})$ and $p(h_j = 1|v) = \sigma(b_j + \sum_i v_i w_{i,j})$, where $\sigma(x) = \frac{1}{1+e^{-x}}$.

In order to find the parameters $\{W, a, b\}$, the contrastive divergence algorithm is applied [9].

The standard configuration of a deep RBM is a pipeline that includes a RBM (it may have multiple stacked RBMs) and, connected to its output a final classifier. The last can be a standard classifier as K-nearest neighbors (KNN), decision trees or multilayer perceptrons (MLP), among others [1,5]. After training the RBM, the outputs from the hidden neurons feed the final classifier.

## 3 Pruning with discriminative measures

Although the standard way is widely accepted, there are not explicit proofs that the last layer provides the more discriminative information to the final classifier, and much less in a deep stacked RBM. Then, it is interesting to evaluate the

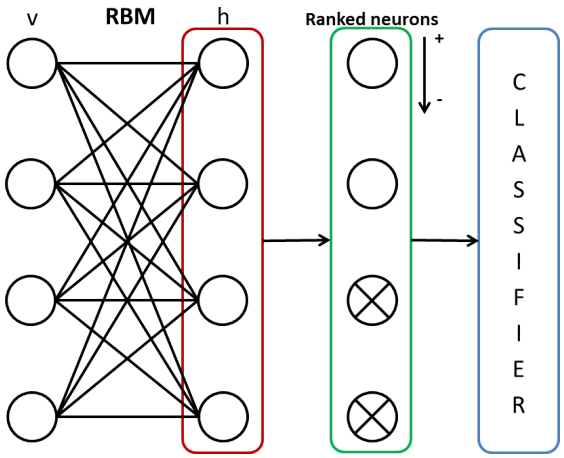



Fig. 2: Proposed DBM.

discriminative capacity of every units, to rank them and to use the best to feed the final classifier (see Figure 2).

After RBM training phase, it is feeding again with the training samples and the activations are collected in every unit for each class. Then, it is possible to think about activation probabilities of the classes in each unit. Consequently, the more different these activations are, the more discriminative that the unit could be. In this work, *information gain* and *Pearson correlation* are proposed to measure this in order to rank the units. The general steps used in the multi–class approach are described in Algorithm 1.

**Algorithm 1** Discriminative evaluation

**Require:** An unsupervised trained RBM
**for all** class **do**
    calculate the propagated value in the hidden layer for each training vector
**end for**

**Require:** The outputs of the RBM (the propagated vectors)
**for all** neuron $i$ **do**
    estimate separately the histograms of the output data for each class.
    calculate $i$'s discriminative value $D_i$ according to the selected measure.
**end for**
Rank the neurons according to their discriminative value in descending order.

**Require:** Ranked neurons
**for** $i \leftarrow 1$ **to** *total number of neurons* **do**
    use the first $i$ neurons to classify the data using Knn.
**end for**

### 3.1 Information gain

The information gain is used to measure about the 'information gained' in the classification task, in presence or absence of a neuron, by the decrease of global entropy. Entropy is considered as a measure of the unpredictability of the system, then, if the randomness of the given variable is known, the amount of information provided by an event can be estimated. For a random variable $X$ with probability mass function $p$, it is computed as $H(X) = -\sum_x p(x) \log_2 p(x)$.

Then, a more probable event is less informative and it is possible to define the information for a particular event as $I(x) = -\log_2 p(x)$, and its expected value over all possible values of $x$ leads to the Shannon's entropy. From Shannon's entropy we can define the conditional entropy of a random variable $X$ given the random variable $Y$ by:

$$H(X|Y) = \sum_{x,y} p(x,y) \log_2 p(x|y) \tag{4}$$

where $p(x, y)$ is the joint probability that $X = x$ and $Y = y$. Using that, the information gain of a class for a unit is defined as:

$$IG\left(Class, Attribute\right) = H\left(Class\right) - H\left(Class|Attribute\right) \tag{5}$$

In this context, the hidden unit is an attribute.

### 3.2 Pearson correlation

In this work, the Pearson correlation coefficient (PCC) [16] is computed as the normalized covariance: 1 means direct correlation, $-1$ means inverse correlation and 0 denoting the absence of any relationship. The idea is that the correlation of

the samples within a class is expected to be greater than the correlation between classes [26], in this way, the ranking of neurons uses the correlation values. This coefficient can be expressed as:

$$r = r_{xy} = \frac{cov\,(x, y)}{\sigma_x \sigma_y} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)}\ \sqrt{(\sum y_i^2 - n\bar{y}^2)}} \tag{6}$$

where $\sigma$ is the standard deviation, $n$ is the sample size and $\bar{x}, \bar{y}$ are the means. When $x$ and $y$ come from the same class, this coefficient is interpreted as the intra–class correlation while, when they come from different classes, as the inter–class correlation. It means, for each neuron, the correlation is obtained using the neuron's output values for the samples of the classes. Then the ranking is carried out as described in Algorithm 1.

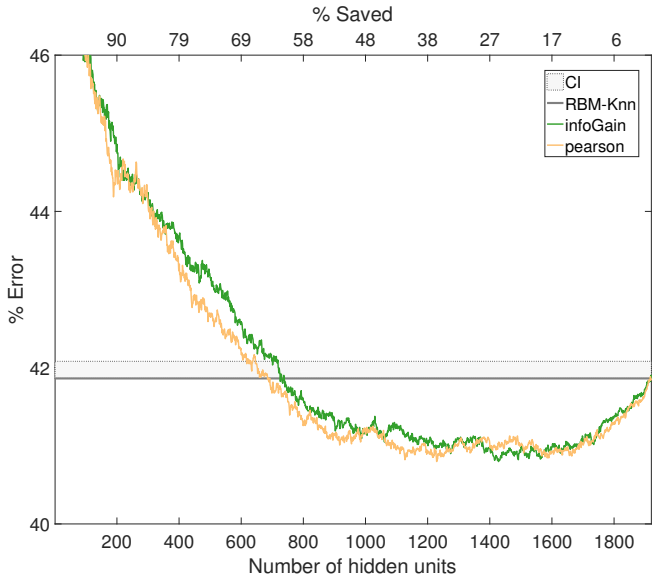
## 4 Experiments and Results

The baseline was defined as a standard DBM using a 10–fold cross-validation (CV)[17], and a confidence interval (CI) with 0.05 level of significance was computed. Then, the pruning performance was evaluated on the final classifier with 10–fold CV. For each pruning, a new final classifier is trained an tested. The experiments were performed using an RBM with 30 visible and 1920 hidden units for the Interface corpus, for the EmoDB database 30 visible and 960 hidden neurons were used. The number of hidden units were set based on the number of audio samples [19]. The baseline is represented with a solid-line and the confidence interval (CI) with a dashed-line in the Fig. 3. The pruning was done with the best 200 units and then, adding 200 successively for Interface, and using 100 and an increment of 100 for EmoDB. When the performance of the pruned networks crosses the baseline CI, it is a good point to stop the searching. However, it is possible to see the better performances reached by the pruned networks.

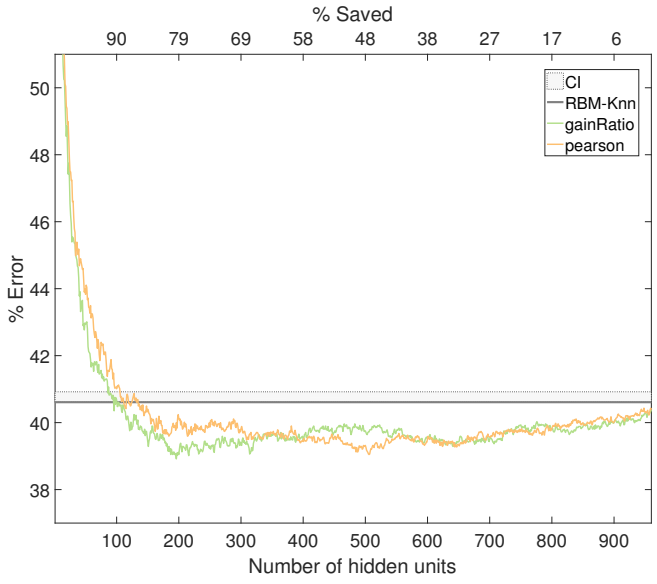
## 5 Discussion

Results show that the two proposed measures are useful to achieve an acceptable error rate with fewer neurons. As can be seen, the pruned networks use less units than the full RBM and reach better classification rates. This may keep the advantages in classification using a big net and to improve the results using a standard DBM.

The results indicate that once a suitable number of initial neurons has been chosen, pruned networks with less than 50% of the neurons produce better-than-baseline error results. For example in the Fig. 3 (a), around 40% of the total neurons are needed to achieve the same performance than the baseline while in (b), only 10% is needed. In both figures, it can be seen that the error decreases until that adding more neurons does not give more information and make the net more complex.




(a)




(b)

Fig. 3: Pruning results using INTERFACE(a) and EmoDB(b) corpora. Test were performed using different RBM configurations.

In the post-training pruning method for restricted Boltzmann machines presented in this work, the hidden units were ranked and then pruned using *information gain* and *Pearson correlation*.

In this work, we used the pruning scheme in multi-class classification and it obtain a good performance and it is very promising to be applied in other tasks. Finally, this can be considered as a method for feature extraction (from the hidden units of a RBM).

In future work, more task will be evaluated and more techniques to measure the discriminative ability of neurons will be explored.

## Acknowledgements

The authors wish to thank to *Universidad Autónoma Metropolitana* from México; *Agencia Nacional de Promoción Científica y Tecnológica* (ANPCyT)(with PICT-2015-977), *Universidad Nacional del Litoral* (with CAID-PJ-50020150100055LI) and *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET), from Argentina, for their support. In addition, they want to thank ELRA for supplying the emotional speech synthesis database, catalogue reference: ELRA-S0329.

## References

1. Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V., Amir, N.: The automatic recognition of emotions in speech. In: Cowie, R., Pelachaud, C., Petta, P. (eds.) Emotion-Oriented Systems, pp. 71–99. Cognitive Technologies, Springer Berlin Heidelberg (2011)
2. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. of 9th European Conference on Speech Communication and Technology (Interspeech). pp. 1517–1520 (Sep 2005)
3. Cao, F., Liu, B., Park, D.S.: Image classification based on effective extreme learning machine. Neurocomputing **102**, 90 – 97 (2013)
4. Castellano, G., Fanelli, A.M., Pelillo, M.: An iterative pruning algorithm for feedforward neural networks. IEEE Transactions on Neural Networks **8**(3), 519–531 (1997)
5. Cen, L., Yu, H.L.Z.L., Dong, M., Chan, P.: Machine learning methods in the application of speech emotion recognition. INTECH Open Access Publisher (2010)
6. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research **11**(Feb), 625–660 (2010)
7. Guo, X.L., Wang, H.Y., Glass, D.H.: A growing bayesian self-organizing map for data clustering. In: 2012 International Conference on Machine Learning and Cybernetics. vol. 2, pp. 708–713 (July 2012)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
9. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation **14**(8), 1771–1800 (2002)

10. Hinton, G.E.: A practical guide to training restricted boltzmann machines. Department of Computer Science, University of Toronto (2010)
11. Hinton, G.E.: A Practical Guide to Training Restricted Boltzmann Machines, pp. 599–619. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
12. Huang, F.J., Boureau, Y.L., LeCun, Y., et al.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
13. Huang, G.B., Wang, D.H., Lan, Y.: Extreme learning machines: a survey. International Journal of Machine Learning and Cybernetics **2**(2), 107–122 (2011)
14. Hussain, S., Alili, A.A.: A pruning approach to optimize synaptic connections and select relevant input parameters for neural network modelling of solar radiation. Applied Soft Computing (2016)
15. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 609–616. ICML '09, ACM, New York, NY, USA (2009)
16. Lee Rodgers, J., Nicewander, W.A.: Thirteen ways to look at the correlation coefficient. The American Statistician **42**(1), 59–66 (1988)
17. Michie, D., Spiegelhalter, D., Taylor, C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood, University College, London (1994)
18. Sánchez-Gutiérrez, M., Albornoz, E.M., Rufiner, H.L., Close, J.G.: Post-training discriminative pruning for rbms. Soft Computing (2017). https://doi.org/10.1007/s00500-017-2784-3
19. Sánchez-Gutiérrez, M.E., Albornoz, E.M., Martinez-Licona, F., Rufiner, H.L., Goddard, J.: Pattern Recognition, chap. Deep Learning for Emotional Speech Recognition, pp. 311–320. Springer International Publishing (2014)
20. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. Proc. Interspeech, ISCA pp. 148–152 (Aug 2013)
21. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.F., Pantic, M.: A survey of multimodal sentiment analysis. Image and Vision Computing **65**, 3 – 14 (2017). https://doi.org/https://doi.org/10.1016/j.imavis.2017.08.003, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing
22. Stanley, K.O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. Evolutionary Computation **10**(2), 99–127 (2002)
23. Sutskever, I., Hinton, G.E.: Learning multilevel distributed representations for high-dimensional sequences. In: AISTATS. vol. 2, pp. 548–555 (2007)
24. Suzuki, K., Horiba, I., Sugie, N.: A simple neural network pruning algorithm with application to filter synthesis. Neural Processing Letters **13**(1), 43–53 (2001)
25. Tao, J., Kang, Y.: Features importance analysis for emotional speech classification. In: Tao, J., Tan, T., Picard, R. (eds.) Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, vol. 3784, pp. 449–457. Springer Berlin Heidelberg (2005)
26. Wei, X., Li, K.C.: Exploring the within-and between-class correlation distributions for tumor classification. Proceedings of the National Academy of Sciences **107**(15), 6737–6742 (2010)