*Journal of Communications Technology, Electronics and Computer Science, Issue 34, 2021*
*ISSN 2457-905X*

# Vector quantization for selecting the number of neurons

Tomas Ruzgas[1], Jurgita Arnastauskaitė[1,2], Laura Kižauskienė[2], Mindaugas Bražėnas[3]
[1]Department of Applied Mathematics, [2]Department of Computer Sciences, [3]Department of Mathematical Modelling
Kaunas University of Technology
Kaunas, Lithuania

***Abstract*— The paper deals with the strategy of selecting the number of neurons based on vector quantization methods. Two methods based on neural networks are analyzed: self-organizing map and neural gas. A method of selecting the number of neurons according to the specifics of the data analyzed is presented.**

***Keywords*— *Vector quantization; number of neurons***

## I. INTRODUCTION

Vector quantization [1] is the process by which *d*-dimensional data set vectors $X(1), X(2), \dots, X(n)$, where *n* denotes the number of vectors, are replaced by a smaller number of *d*-dimensional vectors $Z(1), Z(2), \dots, Z(m)$, $m < n$. Vector quantization techniques are usually used to compress data (such as sound, image). Though, they are suitable for data clustering and classification as well. The latter methods include self-organizing maps (SOM) [2], vector training quantization [3], the neural gas [4] method and others. Quantization error is calculated to evaluate quantization results. This error acquires the least significance when *m* = *n*, but the goal of quantization methods is to reduce *m*. However, determining the value of *m* is usually problematic, seeking for the result to be acceptable for the problem being solved. This paper deals with vector quantization methods based on neural networks - the self-organizing map and the neural gas method. In these methods, the number of quantized vectors *m* is called the number of neurons. The aim of the study is to determine the number of neurons *m* taking into account the specificity of the data being analyzed.

## II. VECTOR QUANTIZATION METHODS

Let the matrix of the data set be denoted as $X = \{X(1), X(2), \dots, X(n)\} = \{x_t^k, t = 1, \dots, n, k = 1, \dots, d\}$, its rows as vectors $X(t) \in R^d$, i.e. $X(t) = (x_t^1, x_t^2, \dots, x_t^d)$, $t = 1, \dots, n$, where $x_t^k$ is the *k*-th component of vector *t*, *d* is the number of components (dimension), *n* is the number of vectors to be analyzed. Both the neural gas method and the self-organizing map form a neuronal array *Z*. Neurons are vectors with a dimension number equal to *d*. In the neural gas method, the neural network is one-dimensional $= \{Z(1), Z(2), \dots, Z(m)\}$, where $Z(\tau) = \{z_\tau^1, z_\tau^2, \dots, z_\tau^d\}$, $\tau = 1, \dots, m$. The SOM network is two-dimensional $Z = \{Z(i,j), i = 1, \dots r, j = 1, \dots, s\}$, where $Z(i,j) = \{z_{ij}^1, z_{ij}^2, \dots, z_{ij}^d\}$, *r* is the number of rows, *s* is the number of columns, then the number of neurons is $m = r \times s$. The aim of the methods is to change the values of the neurons so that they reflect the properties of the analyzed vector $X(t)$, $t = 1, \dots, n$, i.e. at the end of training, neurons become quantized vectors of vectors $X(t)$.

Random start principle generates random initial values of neural components in the interval $(-0.5 \cdot 10^{-5}; 0.5 \cdot 10^{-5})$ for the neural gas or $(0; 1)$ SOM before training the network. During training, the vectors of the training set *X* are sequentially presented to the network for a determined number of times. Each vector is submitted for a *q* number of times in the network. Since the number of analyzed vectors is equal to *n*, the number of training iterations is $h_{max} = q \times n$. By submitting vector $X(l)$, $l \in \{1, \dots, n\}$ to the network, the Euclidean distance is calculated from $X(l)$ to all neurons.

In SOM training, by placing the vector $X(l)$ into the network, the Euclidean distance from it to all neurons in the network is calculated. The winner neuron $\tilde{Z}$ is found with the shortest distance from $X(l)$. The values of the neurons are changed according to the formula:

$$Z_{(h+1)}(i,j) = Z_{(h)}(i,j) + \theta_{(h)}(i,j) \cdot (X(l) - Z_{(h)}(i,j)),$$

where $h$ is the iteration number, $\theta_{(h)}(i,j)$ is called neighborhood function whose value depends on the number of iteration being performed and the position of the recalculated neuron in the network relative to the neuron winner. Convergence of the process requires that $\theta_{(h)}(i,j) \to 0$, when $h \to \infty$.

In the neural gas method, neurons $Z(1), Z(2), \dots, Z(m)$ are replaced by neurons $W(1), W(2), \dots, W(m)$, where $W(\tau) \in \{Z(1), Z(2), \dots, Z(m)\}$, $\tau = 1, \dots, m$, such that

$$\|W(1) + X(l)\| \le \|W(2) + X(l)\| \le \dots \le \|W(m) + X(l)\|$$

Then the distance from $X(l)$ to the first neuron $W(1)$ is the smallest. This neuron is also called the winning neuron. The values of all neurons are changed according to the formula:

$$W_{(h+1)}(\tau) = W_{(h)}(\tau) + E_{(h)} \cdot \theta_\lambda \cdot (X(l) - W_{(h)}(\tau)),$$

where $h$ is the iteration number, $E_{(h)} = E_g(E_f/E_g)^{(h/h_{max})}$, $\theta_\lambda = e^{-(\tau-1)/\lambda_{(h)}}$, $\lambda_{(h)} = \lambda_g(\lambda_f/\lambda_g)^{(h/h_{max})}$, the values of the parameters $\lambda_g$, $\lambda_f$, $E_g$, $E_f$ are selected before the network training.

Once the network is trained, its quality must be evaluated. In vector quantization methods, the quantization error is usually estimated by the formula

$\Delta = \frac{1}{n}\sum_{l=1}^{n}\|X(l) - \tilde{Z}\|$,
where $\tilde{Z}$ is the winner of vector $X(l)$. In the neural gas method $\tilde{Z} = W(1)$.

## III. Empirical Study

The vector quantization methods discussed above have been applied to real world data used by many researchers in their work:

- Fisher [5] used *iris data*. The lengths and widths of the three species of iris, sepals and petals have been measured. 4-dimensional vectors were created, *d = 4*, *n = 150*.
- Quinlan [6] used *data for cars manufactured in the USA, Europe and Japan*. Car fuel consumption, number of cylinders, engine displacement, horsepower, weight, speed, year of manufacture have been indicated. 7-dimensional vectors were created, *d = 7*, *n = 398*.
- Asuncion & Newman [7, 8] used *wheat data*. Five wheat species were analyzed. 12 parameters were measured using a digital photo of each grain: grain area, perimeter, color characteristics, etc. 12-dimensional vectors were created, *d = 12*, *n = 400*.

The results of the neural gas method depend on the training parameters $\lambda_g$, $\lambda_f$, $E_g$, $E_f$, the number of training steps $q$ and the number of neurons $m$. There is no theoretical evidence of solutions' convergence, thus in order to find the most suitable solution the parameters have to be selected empirically. Alhoniemi [9] has found that the smallest quantization errors are obtained when $\lambda_f = 0{,}01$, $E_f = 0{,}1$, $\lambda_g = m/2$, $E_g = 0.5$. It has also been found that sufficiently stable results are obtained at q = 200. Increasing this number is not important as the quantization error decreases insignificantly.

In the neural gas and SOM methods, the smallest quantization error (Δ) is obtained when the number of neurons is $m = n$ and $n$ is the number of vectors to be analyzed. However, the deviation from a given value of $m = m'$ is insignificant, compared to the smallest. The graphs of quantization error variations for three different dimensions' data sets are presented in Figure 1. The lowest quantization error is obtained by analyzing the iris data (*d = 4*, *n = 150*), the higher – by analyzing cars (*d = 7*, *n = 398*), and the highest – wheat data (*d = 12*, *n = 400*). In addition, it has been observed that the scatter graph of the errors follow exponential function $u = av^b$ with a negative degree index $b$ <*0*. Determination coefficient of approximation for $R^2 > 0.98$ (SOM) and $R^2 > 0.84$ for the neural gas method.

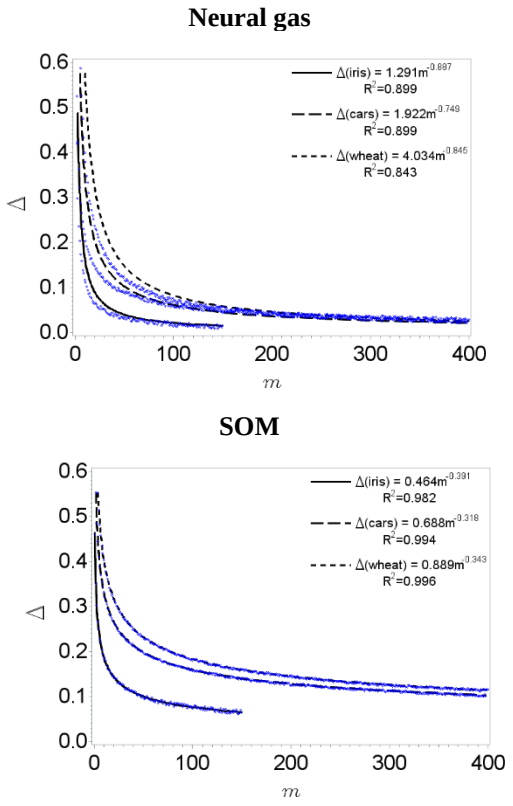


Fig. 1. Quantization errors

In the study of the dependence of the number of neurons on the number of vectors in dimension *d*, the generated data was used. 10000 sets of $n$ vectors were generated with the dimension of the vectors being *d*. SOM and neural gas networks were trained with these sets. The Simulation study was performed with small and medium quantities of vectors (20, 50, 100, 200) and neuron numbers $m = 4, \ldots, n$. The values of the number of neurons $m' < m$ were recorded under the condition when the quantization error value of this number of neurons differs from the smallest ($m = n$) by not more than $\varepsilon = 0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.2$. The percentage of the number of neurons $m'$ from the total number of neurons $m = n$ was calculated.

The dependence of the number of neurons $m'$ on the dimension number *d* of the analyzed vectors is shown in Figures 2 and 3. A higher percentage means that the network needs to be composed from more neurons in order to obtain the desired accuracy error. Figures 2 and 3 denote that the increase of *d* increases the percentage of neurons with a fixed error $\varepsilon$. This means that the higher the *d*, the more neurons need to be sampled to get good results in terms of quantization error. Comparison of the results of the neural gas method with the SOM network showed that the SOM method allows the rejection of a significant proportion of neurons without significant loss of accuracy. For example, with $n = 50$, $d = 20$, $\varepsilon = 0.2$, only slightly more than 20% of the neurons are sufficient in the SOM method, while the neural gas method

requires nearly 80% of the neurons to achieve the same accuracy.

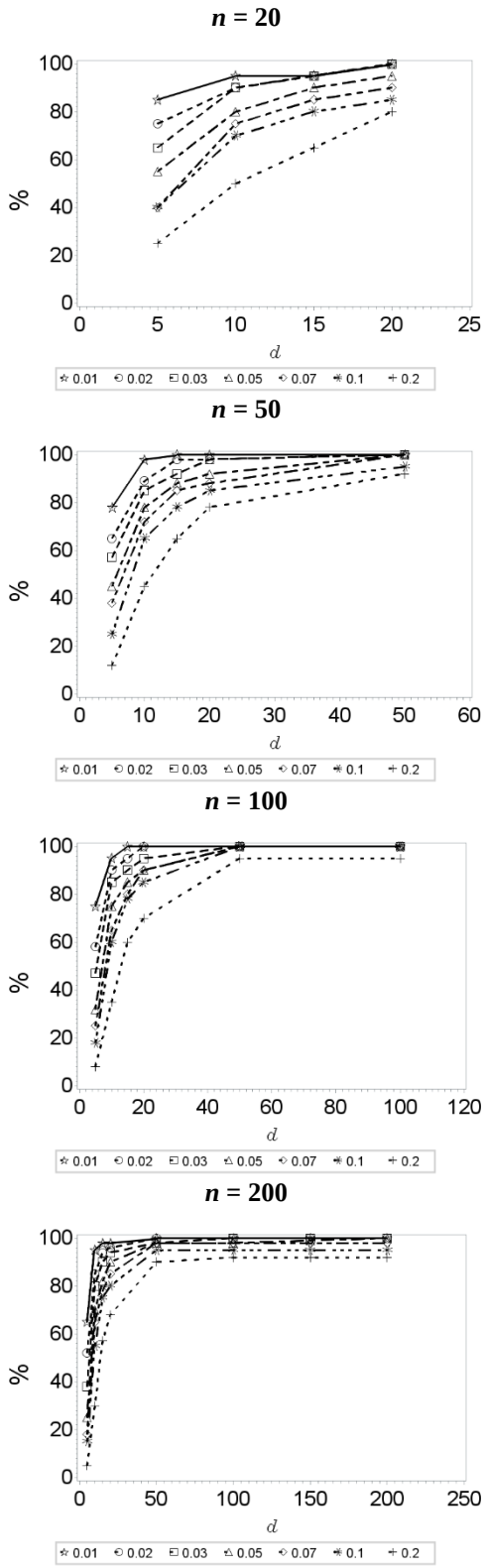


Fig. 2. Number of neurons in the neural gas method

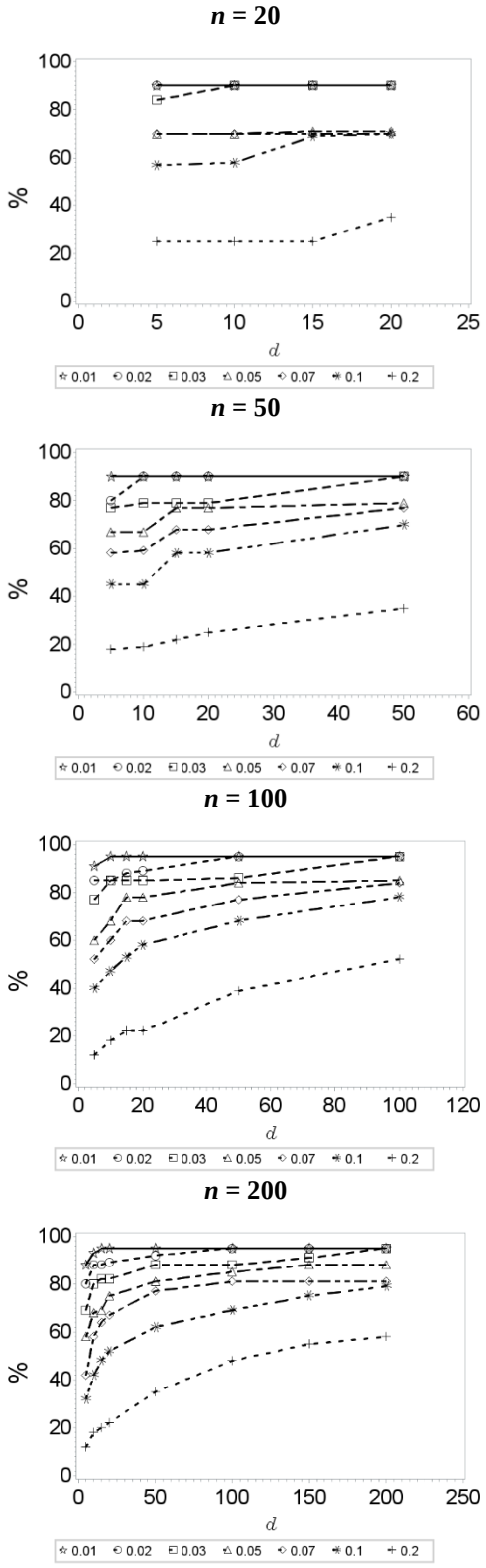


Fig. 3. Number of neurons in the SOM method

## CONCLUSIONS

This paper has investigated two methods for the selecting the number of neurons in two vector quantization methods – self organizing map and neural gas method. A network consisting of several neurons has been studied to reduce the set of analyzed vectors so that the quantization error would be small in magnitude from the smallest error obtained when the number of neurons coincides with the number of vectors used. The findings of the study indicate that with a larger number of vector dimensions, the network needs to be built up with more neurons so that quantized vectors can reflect the properties of the analyzed vectors as accurately as possible. The obtained results indicate that compared to the neural gas method, a SOM network can reach the same accuracy with a significantly smaller number of neurons.

## REFERENCES

[1] Gray, R.M. (1984). Vector quantization, IEEE ASSP Mag., pp. 4-29.

[2] Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21 (1–3), pp. 1-6

[3] Wu, Z., Yu, J. (2019). Vector quantization: a review. *Frontiers of Information Technology & Electronic Engineering*. 20, pp. 507–524.

[4] Hammer, B., Strickert, M. & Villmann, T. (2005). Supervised Neural Gas with General Similarity Measure. *Neural Processing Letters*. 21, pp. 21–44

[5] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7, pp.179–188.

[6] Quinlan, R. (1993). Combining Instance-Based and Model-Based Learning. *In Proceedings on the Tenth International Conference of Machine Learning*, pp. 236-243

[7] Asuncion, A., Newman, D.J. (2019). UCI *Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA (http://www.ics.uci.edu/˜mlearn/MLRepository.html)

[8] Ajaz R.H., Hussain L. (2015), Seed Classification using Machine Learning Techniques, *Journal of Multidisciplinary Engineering Science and Technology*, Vol. 2, No. 5, pp. 1098–1102.

[9] Alhoniemi, E., J. Himberg, J., Parhankangas, J., Vesanto, J. (2000). *SOM Toolbox for Matlab* 5, Helsinky University of Technology, Report A57.