Individual projects

A. On the quantization error in SOM vs. VQ: A critical and systematic study

Teuvo Kohonen, Ilari T. Nieminen, and Timo Honkela

The self-organizing map (SOM) is related to the classical vector quantization (VQ). Like in the VQ, the SOM represents a distribution of input data vectors using a finite set of models. In both methods, the quantization error (QE) of an input vector can be expressed, e.g., as the Euclidean norm of the difference of the input vector and the best-matching model.

Some attempts have been made to compare the quantization errors in the SOM vs. the same errors in VQ. It is usually taken as self-evident that if, for instance, the models or "codebook vectors" are optimized in the VQ so that the sum of the squared QEs is minimized for given training vectors, it will be impossible to find any other set of models that produces a smaller *rms QE* (square root of the mean square of QE over independent test data). Thus the rms QE also in the SOM is supposed to be larger. Therefore it has come as a surprise that *the rms QE of the SOM may sometimes be smaller than that of the VQ* (cf., e.g., [1] and [6]).

We have found out that *this effect depends most strongly on the ratio of the number of training vectors and the number of model vectors.* If this ratio is *small, on the order of small integers*, the rms QE of the SOM is usually smaller than that of the VQ. However, *the training vectors must also have a significant local variance in sufficiently many dimensions.*

Artificial data

The first experiment was carried out with artificially generated, normally distributed random data with zero mean and identity covariance, and a 10x14 SOM. The ratio of the QEs in the SOM and the VQ with 140 codebook vectors is displayed as a function of the number of training vectors per model. Fig. I.4 represents the results. With the input dimensionality 10 the rms QE of the VQ was always smaller. For the dimensionalities 20 and 50 , the "break even" point (where the rms QEs of the SOM and the VQ are equal) occurred at the argument value 12.2 Below this point, the rms QE was always smaller in the SOM. The lower limit of the input dimensionalities for this effect to occur seems to exist between 10 and 20.

ISOLET data

The input vectors in this experiment consisted of 617 acoustic features extracted from spoken letters of the English alphabet [2]. In order to achieve a sufficient statistical accuracy, the *repeated holdout validation* was used. In it, a number of training samples is picked up at random from the available data set, while the rest of it is set aside for testing.

In the ISOLET experiment we had 7797 input vectors available. For instance, with the SOM array size 10x14 = 140 and the argument value 50, we selected M = 140x50 = 7000 samples at random from the basic data set for the construction of one SOM/VQ pair, while the rest was set aside for testing. At lower argument values, less data are needed for the construction of a SOM/VQ pair, whereupon more data can be reserved for testing. This random selection of the training vectors was repeated 20 times for every argument value, and a new SOM/VQ pair was constructed every time. The averages over the repeated evaluations of the rms QEs were then formed for every argument value.

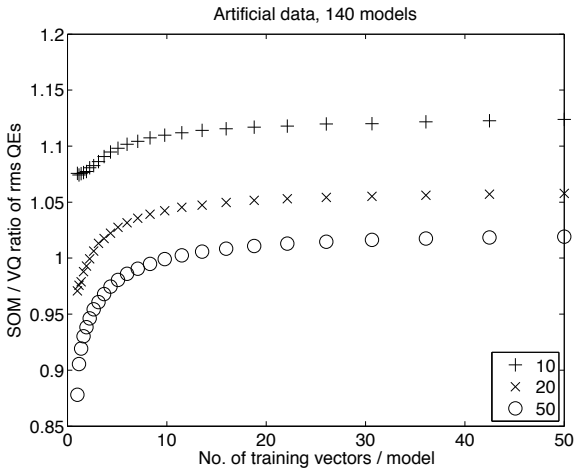

Figure I.4: Ratio of the rms QEs in the SOM and the VQ for the artificially generated random-data set, as a function of the number of training vectors per model, and for the dimensionalities 10, 20, and 50, respectively. The number of models was 140.

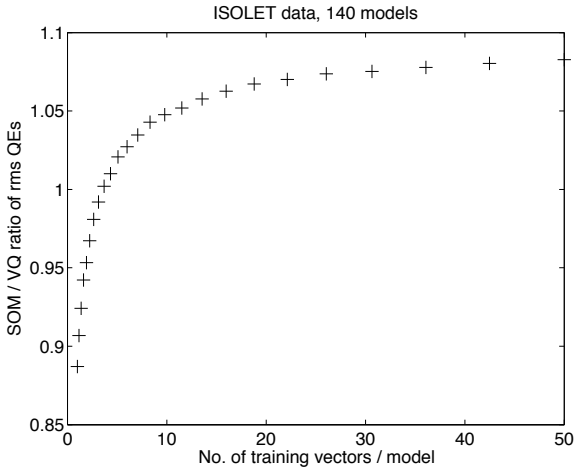

Figure I.5: Ratio of the rms QEs in the SOM and the VQ for the ISOLET data set, as a function of the number of training vectors per model and for 140 models.

In Fig. I.5 we display the ratio of the rms QEs in the SOM and in the VQ as a function of the number of training vectors per model. It can be seen that the "break even" point (at which the rms QEs in the SOM and the VQ are equal) is about 3.7. It is to be noted that the "effective dimensionality" or "fractal dimension" of real data is always much less than the true input dimensionality.

Reuters data

Our third experiment was based on the text corpus collected by the Reuters Corp. No original documents were available to us, but Lewis et al. [4] have prepared a test data set on the basis of this corpus for benchmarking purposes, preprocessing the textual data, removing the stop words, and reducing the words into their stems. The input vectors, with the true dimensionality of 233, were formed as weighted word histograms.

The general arrangement of this experiment was similar to that with the ISOLET data.

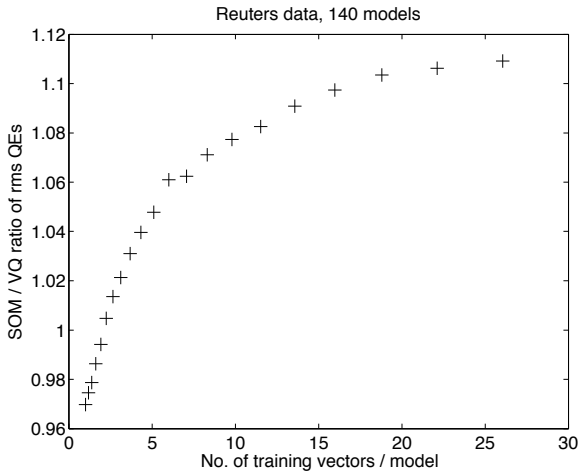

Figure I.6: Ratio of the rms QEs in the SOM and the VQ for the Reuters data set, as a function of the number of training vectors per model and for 140 models.

This time we had only 4000 input samples (documents) available. The averaged results are shown in Fig. I.6 for the SOM array size 10x14. The "break even" point was about 2.45.

Discussion

An explanation of the observed effect seems to ensue from statistics. Each model vector in the VQ is determined as the average of those training vectors that are mapped into the same Voronoi domain as the model vector. On the contrary, each model vector of the SOM is determined as a weighted average of all of those training vectors that are mapped into the "topological" neighborhood around the corresponding model. The number of training vectors mapped into the neighborhood of a SOM model is generally much larger than that mapped into a Voronoi domain around a model in the VQ. Since the SOM model vectors are then determined with a significantly higher statistical accuracy, the Voronoi domains of the SOM are significantly more regular, and the resulting rms QE may then be smaller than in the VQ. For a more detailed discussion, see [3].

References

[1] O. Bação, V. Lobo, and M. Painho, "Self-organizing maps as substitutes for k-means clustering," in *Computational Science - ICCS 2005, Lecture Notes in Computational Science,* Berlin, Heidelberg, Germany: Springer-Verlag, 2005, pp. 476-483.

[2] R. A. Cole, Y. Muthusamy, and M. A. Fanty, *The ISOLET Spoken Letter Database,* Technical Report 90-004, Computer Science Department, Oregon Graduate Institute, 1994.

[3] T. Kohonen, I. T. Nieminen, and T. Honkela, "On the Quantization Error in SOM vs. VQ:A Critical and Systematic Study," in *Advances in Self-Organizing Maps, Lecture Notes in Computational Science LNCS-5629,* Berlin, Heidelberg, Germany: Springer-Verlag, 2009, pp. 133-144.

[4] D.D. Lewis, Y. Yang, T.G. Rose, and T. Li, "RCV1: A new benchmark collection for text categorization research," *J. Machine Learning Research*. Vol. 5, pp. 361-397, 2004.

[5] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing,* Cambridge, MA: MIT Press, 1999.

[6] J. D. McAuliffe, L. E. Atlas, and C. Rivera, "A comparison of the LBG algorithm and Kohonen neural network paradigm for image vector quantization," in *Proc. ICASSP-90, Acoustics, Speech and Signal Processing,* Vol. IV, Piscataway, N.J.: IEEE Service Center, 1990, pp. 2293-2296.