Vector Quantization with Growing Cell Structures Applied in Speaker Verification

Bogdan Sabac and Inge Gavat
University Politehnica from Bucharest
Aleea Faurei, 8/11, cod 78409, Bucharest, Romania
Phone: +401-6673831
email:{sbogdan, inge}@helix.elia.pub.ro

ABSTRACT: We present a self-organising neural network which performs unsupervised learning and can be used for vector quantization. The main advantage over existing approaches, e.g., the Kohonen feature map, is the ability of the model to automatically find a suitable network structure and size. This is achieved through a controlled growth process which also includes occasional removal of units. The vector quantization algorithm can be used to construct a classification system for speaker recognition. For each registered speaker we build two codebooks models, respectively for MFCC (mel frequency cepstral coefficients) and DMFCC (delta mel frequency cepstral coefficients). The speaker verification is done separate by phone (MFCC) modelling and first order temporal phone transition (DMFCC) modelling. The algorithm is evaluated on a database that includes 25 speakers each of them recorded in 12 different sessions. The overall performance was 99.5%. That is, in 99.5% of the trials, the right speaker was correctly accepted and the impostor speaker correctly rejected.

KEYWORDS: vector quantization, speaker verification, growing cell structures, confusion matrix.

INTRODUCTION

Self-organizing neural network models generate mappings from high-dimensional signal spaces to lower-dimensional topological structures. These mappings are able to preserve neighborhood relations in the input data and have the property to represent regions of high signal density on correspondingly large parts of the topological structure. This makes them interesting for applications in various areas ranging from speech recognition and data compression to combinatorial optimization. The fact that similar mappings can be found at various places in the brains of humans and animals indicates that preservation of topology is an important principle at least in natural-signal processing systems. It has been noted that the predetermined structure and size of Kohonen's model imply limitations on the resulting mappings. A number of variations have been proposed concerning networks with variable topology or variable number of elements. The network presented in this contribution has a flexible as well as compact structure, a variable number of elements, and a free-dimensional topology. Recently it was demonstrated that the new model improves over Kohonen's feature map with respect to various important criteria Fritzke (1993a). We acknowledge, however, that the our model is a adaptation at the Fritzke's "growing cell structures" model and it is an extension of his work rather than a completely different formalism. First we will present the network for vector quantization and then the speaker verification experiment.

UNSUPERVISED GROWING CELL STRUCTURES

PROBLEM DEFINITION

Before we describe our network model, it seems appropriate to exactly define the kind of problems the network is supposed to solve. In the first place, we have a number of n-dimensional input signals obeying an unknown probability distribution P(x). With $V = R^n$ we denote the vector space the input signals stem for.
Our objective is to generate a mapping from V onto a discrete free-dimensional topological structure (Fritzke (1993b) use a mapping onto a discrete k-dimensional topological structure with a predefined size for k). The mapping should have the following proprieties:

- Similar input signals are mapped onto topologically close elements of A
- Topologically close elements of A should have similar signals being mapped onto them
- Regions of V where the probability density of the input vector distributions high should be represented by correspondingly many elements in A.

The first two points mean that the mapping should preserve similarity relations in forward and backward direction. If the dimensionality of A is smaller than that of V, a dimensionality reduction is performed. If it is in spite of that possible to preserve the similarity relations, then the complexity of the data is reduced without loss of information. The third point means that we gain some information about the unknown probability density of the input signals.

NETWORK ARCHIECTURE AND DYNAMICS

The initial topology of the network A is composed from two cells (or neurons) connected by a edge. The edges denote topological neighborhood relations. During a self-organization process described further below new cells will be added to the network and superfluous cells will be removed.

Every cell c has an n-dimensional synaptic vector w_c attached. This vector may be seen as the position of c in the input vector space. We denote with w the set of all synaptic vectors $w_i \subset A$. A mapping g(x) from the input vector space V onto the network A can now be defined by mapping every input signal to the cell with the nearest position (or reference vector). More formally we write

$$g_w : V \to A, \quad if \ x \subset V \quad then \ g_w(x) \subset A \tag{1}$$

with g_w the so called best-matching unit (bmu) being defined through

$$\| w_{bmu} - x \| = \min \| w - x \| \tag{2}$$

where $\| \ \|$ denotes the Euclidean vector norm.

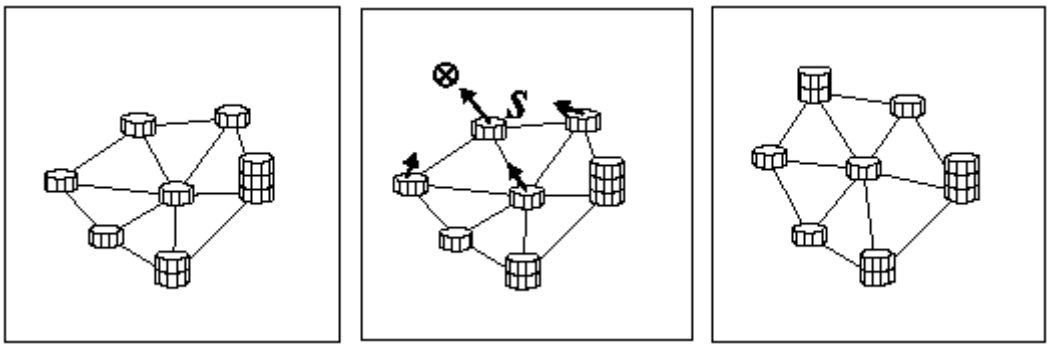
In principle the adaptation of the synaptic vectors in our model is done as earlier proposed by Kohonen (1982):

i). Determine the best-matching unit for the current input signal.

ii). Increase matching at the best matching unit and its topological neighbors.

In Kohonen's model the strength of the adaptation is decreasing according to a cooling schedule. Moreover, the topological neighborhood inside which significant changes are made is chosen large at the beginning and decreases then, too. The Fritzke's growing cell structures model follows the same basic strategy. There are, however, two important differences: - the adaptation strength is constant over time. Specifically are used constant adaptation parameters e_f for the best matching unit and e_n for the neighboring cells, respectively.

- only the best-matching unit and its direct topological neighbors are adapted.



a) Initial situation b) Occurrence of a input vector c) After adaptation

Figure 1: One adaptation step for a two-dimensional cell structure.

Only the best-matching unit and its direct neighbors are adapted. The columns represent accumulated error values.

After a adaptation step two important events take place, first at the error accumulated by the first bmu is added the quantity $\| fbmu - x \|$ and second, the age of all edges emanating from fbmu is incremented by 1 with the exception of the age of the edge which connects the fbmu with second bmu neighbour that is set to zero.

Insertion of a new cell take place if after a number of adaptation steps the maximum accumulated error exceeds a insertion threshold. The new cell r is inserted between direct neighbouring cells f and q with f having the largest accumulated error over the insertion threshold and q being a direct neighbour of f with the maximum accumulated error.

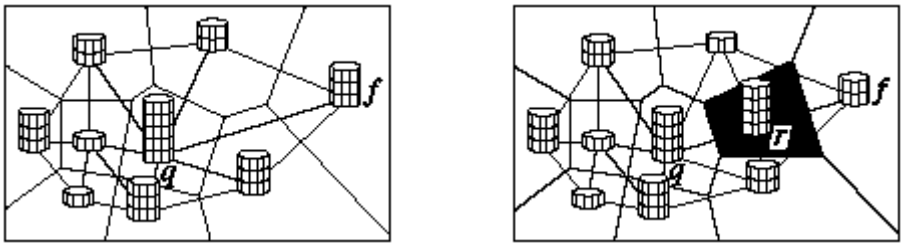


Figure 2: Insertion of a new cell.

If the age of an edge exceeds a predefined age-max then that edge is erased and so are the cells that have no emanating edges.

Start with a two cell structure
repeat
for i:=1 to (all vectors from the training lot) do
1. Find fbmu
2. Find sbmu
3. Increment the age of all edges emanating from fbmu
4. Increment the accumulated error of fbmu
5. Set the age of fbmu-sbmu edge to zero
6. Eliminate all edges with ages over age-max
7. Eliminate cells with no emanating edges
8. Adaptation step for fbmu and his direct neighbours
9. Test the insertion condition
i. find f with the biggest accumulated error
ii. find q a direst neighbour of f with biggest accumulated error
iii. insert the new cell r between f and q proportional with the values of accumulated errors
iv. decrease the accumulated errors of f and q by multiplication with a constant alpha < 1
v. initialise the accumulated error of r
vii. mark the fbmu cell to avoid deletion in step 11
10. Multiply all accumulated errors with a constant delta < 1
11. At every EliminationRate epochs erase the cells that are not marked as fbmu
until no changes in the network structure take places

Figure 3: The vector quantization with growing cell structures algorithm

SPEAKER VERIFICATION SETUP

Speech signal was sampled at 16 kHz with a 8 bit digitizer. The speech signals are analysed with a 30 ms window shifted every 15 ms in order to extract the following parameters from each frame of speech:

(a) 20 mel frequency cepstral coefficients (MFCC). The 0.1-5 kHz frequency range was divided into 64 overlapping equal bands distributed on the Mel scale.

If we denote the output energy of the k-th. filter by $\tilde{Y}(k)$, the mel-warped cepstrum $c_{mel}(n)$ is obtained by taking the shifted discrete cosine transform (DCT) of the Mel-frequency scale (3):

$$c_{mel}(n) = \sum_{k=1}^{N_{bc}} \log\left(\tilde{Y}(k)\right) \cdot \cos\left(n \cdot \left(k - \frac{1}{2}\right) \cdot \frac{p}{N_{bc}}\right) \qquad (3)$$

where:

n=1,2,…,L, is the desired length of the cepstrum

k=1,2,…,N_{bc}, is the number of filters.

(b) 20 delta mel frequency cepstral coefficients (DMFCC) calculated as polynomial expansion over speech segments of five frames in length. Since the spectral transition play an important role in human perception as demonstrated by Furui (1984) the introduction of such features along with the static ones will improve the recognition performance of the system.

With the extracted feature vectors from a speaker, using the growing cell structures algorithm two codebooks are constructed for that speaker. This process is repeated for all speakers in the population. After the centroids are set for

each of them we compute the variance for each dimension. For speaker verification, the feature vectors from a test utterance are applied to the speaker codebook. For that codebook, the centroid that is closest to the test vector is found and the distance to this centroid is accumulated. The speaker corresponding to the codebook is accepted or rejected according to the accumulated distance. The likelihood speaker score is computed according to equation 4:

$$score = score_{MFCC} + score_{DMFCC} =$$

$$= \frac{1}{N_{MFCC}} \sum_{i=1}^{N_{MFCC}} \exp\left(-\sum_{j=1}^{20} \left(\frac{m_j - x_j}{s_j} \right)^2 \right) + \frac{1}{N_{DMFCC}} \sum_{i=1}^{N_{DMFCC}} \exp\left(-\sum_{j=1}^{20} \left(\frac{m_j - x_j}{s_j} \right)^2 \right) \quad (4)$$

where: N = number of vectors MFCC or DMFCC m = mean s = variance x = input vector

The algorithm is evaluated on a database that includes 25 speakers each of them recorded in 12 different sessions. All 25 speakers spoke the same phrase for 120 times, the phrases from recording sessions 1 to 6 are used for training and the phrases recorded in sessions 7 to 12 for testing. The overall performance of the system was 99.5%. That is, in 99.5% of the trials, the right speaker was correctly accepted and the impostor speaker correctly rejected.

Here is a confusion matrix showing the results of another experiment:

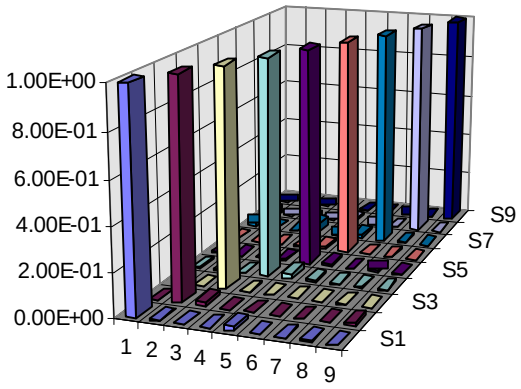


Figure 4: Confusion matrix from a live test with 9 out of 25 speakers.

The scores on the diagonal are the "true speaker" scores. The rest are "impostor" scores. The difference shows how well the system separates them.

CONCLUSIONS

We presented a vector quantization method which incrementally builds up a codebook through interpolation. The growing cell structures method produces good results for a speaker verification problem. The incremental character of the method also has the important advantage that the size of the codebook can be increased until the quantization error falls below a given bound. For the other methods a complete restart with a larger codebook would be necessary if the final error was still above that bound. The presented network is a variant of a more general class of algorithms the underlying principle of which is a growth process controlled by some insertion criterion. In this paper the insertion criterion has been based on the quantization error.

REFERENCES

Fritzke, B. 1993a, "Kohonen feature maps and growing cell structures - a performance comparison", in Advances in Neural Information Processing 5, L. Giles, S. Hanson & J. Cowan, eds., Morgan Kaufmann Publishers, San Mateo.

Fritzke, B. 1993b, "Vector quantization with a growing and splitting elastic net", Proc. of ICANN-93, Amsterdam.

Furui S. 1984, "On the role of dynamic characteristics of speech spectra for syllable perception", Fall Meeting of Acoust. Soc. Japan, 1-1-2: October.

Kohonen, T. 1982, "Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics, 43, pp. 59-69.