Abstract

In this contribution we present two experimental scenarios in which we employed Self Organizing Maps (SOMs) to detect affective states. The first scenario is related towards designing a “Personal Stress Prevention Assistant”: we summarize our efforts to detect affective information related to stress in the posture channel. We show that a person-independent discrimination of stress from cognitive load is feasible when using data from a pressure mat mounted on a seat. The second scenario is embedded towards assisting patients with manic depression: we present preliminary results in detecting emotions from voice data. Our findings illustrate that a person-dependent discrimination of emotions from voice data seems feasible and that a general model might be appropriate to discriminate high and low levels of arousal.

1 Introduction

In recent years several research groups have employed pattern recognition methods in order to automatically detect different affective states of a subject. Applications which would benefit from emotion recognition are predominantly found in the field of Human-Computer Interaction (HCI). By considering affective states of a user, HCI shall become more natural [8]. Modalities which have been used to detect affective states include facial expression [2], speech [7] and physiological signals [4].

In this contribution we show in two scenarios how Self Organizing Maps (SOMs) can be employed to detect affective states. In the first scenario we summarize our efforts to detect affective information related to stress in the posture channel. This scenario goes towards designing a “Personal Stress Prevention Assistant” which could advert us of stressful situations. In the second scenario we present preliminary results towards detecting emotions from voice data. This scenario is embedded in the European research project MONARCA which aims at assisting patients with manic depression. One aspect of MONARCA is to investigate how affective state detection from voice can support the diagnosis and treatment of manic depressive patients.
In the following we first present our experimental designs to collect data and extract features for both scenarios in Section 2. Next we briefly introduce the basic principles of Self Organizing Maps in Section 3. Finally we present and discuss our results in Sections 4 and 5.

2 Data Collection and Feature Extraction

2.1 Mental Stress Scenario

A number of 33 male, healthy subjects were recruited for participating in two sessions of two hours respectively. In one session (the stress condition) the subjects were confronted with high cognitive load under time pressure and social evaluative threat. In another session (the control condition) the subjects were exposed to mild cognitive load without any time pressure or social threat. All subjects were told the cover story that they are taking part in a cognitive experiment. In order to eliminate possible habituation effects, half of the subjects were exposed to the stress condition during the first session and 2 weeks later to the control condition in the second session while for the other half the sequence was vice versa.

In both conditions cognitive load was induced during 20 minutes by the The Montreal Imaging Stress Task (MIST; [3]). The MIST is a computer application which asks the subject to provide correct solutions for mental arithmetic tasks, e.g. what is the result of $57/3 - 7 - 5$? In the stress condition the MIST was configured to induce high cognitive load: the program adapted the level of difficulty and the time limits for each task in a way that each subject could only solve 45-50% of the tasks correctly. In order to induce social evaluative threat, a color bar indicated a comparison between the subject’s performance and the performance of a simulated norm collective. Most of the time the color bar indicated that the subject’s performance was clearly lower than the performance of the norm collective. In order to increase the social threat, the subjects received negative feedback by the study leader. For example, the study leader asked the subject why he is not able to solve such easy arithmetic tasks in time as normal people do. For the control condition the MIST application was configured to induce mild cognitive load: the subjects could solve the arithmetic tasks without any time pressure and without social evaluation. The study leader asked friendly, neutral questions.

During both sessions several signals have been recorded including physiological signals, acceleration and sitting pressure. The left side of Figure 1 shows a subject equipped with wearable sensor systems to measure the signals. On the right side the MIST screen is shown. Based on the experimental data, in recent work we have investigated to what extend affective information related to stress can be found in the electro-dermal activity [9] and in the posture channel [1]. In this contribution we summarize our efforts to discriminate between stress and cognitive load using data from a pressure mat mounted on the seat. More details can be found in [1]. In order to record the pressure

Figure 1: Right: subject equipped with wearable sensor systems: (1, 2, 5, and 6) acceleration sensors, (3) electro-dermal activity sensor, (4) pressure mat. Left: MIST screen during stress condition.
distribution we employed the CONFORMat pressure mat developed by Tekscan. The mat consists of 1024 sensor elements and records the pressure data with a sampling frequency of 25 Hz. For each frame of the pressure mat we computed the center of pressure (CoP) based on the 1024 sensor elements. In Fig. 2 two exemplary frames including the CoP for two sitting postures are shown. In order to characterize the movement properties of the CoP we first computed the corresponding frequency spectra. In a second step, we divided the spectra into 20 frequency bands of equal width and computed the mean value of each band. As a result we obtained a 20 dimensional feature vector for each subject and condition respectively.

2.2 Emotion Recognition Scenario

A number of 10 healthy subjects (5 female) were recruited for the emotion recognition scenario. They had to express 6 emotions by pronouncing two standard sentences twice. The following emotions had to be expressed: disgust, happiness, cold anger, boredom, pride and desperation. The two standard sentences were composed of phonemes from several Indo-European languages:

- “Hat sundig pron you venzy”
- “Fee gott laish jonkill gosterr”

In the beginning of the experiment the subjects were instructed how to express the emotions. For each emotion, a recording of a professional actor who expressed the emotion by speaking the two standard sentences was played. Directly after listening the subject had to speak the same standard sentences twice. The audio recording was made with a standard headset microphone. From the audio recordings a set of 25 utterance features were extracted as summarized in Table 1.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Pitch frequency</td>
</tr>
<tr>
<td>6</td>
<td>Intensity</td>
</tr>
<tr>
<td>5</td>
<td>Amplitude</td>
</tr>
<tr>
<td>2</td>
<td>Voiced vs. unvoiced</td>
</tr>
</tbody>
</table>

Table 1: Audio features
3 Methods

In the following we briefly introduce the basic principle of SOMs and XY-fused Kohonen Networks. Originally, the main purpose of SOMs was to visualize a high-dimensional signal space on a two-dimensional grid of nodes while preserving the topological relationships of the signal space on the two-dimensional display [5].

A SOM consists of a set of unconnected units which are spatially ordered in a two-dimensional grid. Each unit in the map is equipped with a weight vector which has the same dimension as the feature space. In order to map the feature space to the SOM we assume that we have a feature vector $x_{\text{stim}} = [\xi_1, \xi_2, \ldots, \xi_n]^T \in \mathbb{R}^n$. Each unit $S$ in the SOM is equipped with a weight vector $w_s = [\omega_s^1, \omega_s^2, \ldots, \omega_s^n]^T \in \mathbb{R}^n$. The image of the feature vector $x_{\text{stim}}$ on the SOM array is defined as the array element $s$ that matches best with $x_{\text{stim}}$ using a similarity measure $d$:

$$s = \arg\min_i d(x_{\text{stim}}, w_i)$$

In order to preserve the topological relationships of the signal space on the two-dimensional SOM grid, the weight vectors $w_s$ are adapted in a training stage. The basic principle is that SOM units that are close in the grid will activate each other to "learn" from the same input $x_{\text{stim}}$. During the training of the SOM the input vectors are presented to the map in a random order. The unit in the map possessing the weight vector most similar to the presented input is assigned to be the winner. Given an input vector $x_{\text{stim}}$ at time $t$, the update of the weight vector $w_i$ of the node $i$ is then done by

$$w_i(t+1) = w_i(t) + h_{si}(t) [x_{\text{stim}}(t) - w_i(t)]$$

using the neighborhood function

$$h_{si}(t) = \alpha(t) \cdot \exp\left(-\frac{||r_s - r_i||^2}{2\sigma^2(t)}\right)$$

with a learning rate $0 < \alpha(t) < 1$ and with location vectors $r_s, r_i \in \mathbb{R}^2$ of the winning nodes $s$ and the node $i$. Thus the weights of the winning unit and its neighbours become slightly more similar to the presented input vector $x_{\text{stim}}$. As a consequence the units become specialized to those inputs which are frequently mapped onto it. After training, the prediction of new input vectors is done by unit-wise averaging of the outputs associated with the mapped input objects.

In recent years, supervised variants of SOMs were developed which incorporate the class information of the input objects into the learning process [10]. In this work we will focus on the XY-fused Kohonen Network which uses an additional grid of nodes (Ymap) to map the class information [6]. During the training a "fused" similarity measure is used which is based on a weighted sum of similarities between an input vector and all units in the Xmap, and similarities between the corresponding output vector and the units in the Ymap. The common winning unit is determined by the location of the minimum in the fused similarity measure. Both maps are updated simultaneously according to the standard SOM formalism: for the update of the Xmap, the input vector is used while for the update of the Ymap the corresponding output vector is used. The procedure for predicting the class membership of new inputs starts with presenting a new input vector to the network. The position of the winning unit in the Xmap is used to look up the class membership of the corresponding unit in the Ymap: the maximum value of this unit’s weight vector determines the actual class membership.
Table 2: Classification accuracies for leave-one-person-out cross validation of SOMs and XY-fused Kohonen Networks [1].

<table>
<thead>
<tr>
<th>Method</th>
<th>Grid size</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td>3x3</td>
<td>61.25 ± 3.48</td>
</tr>
<tr>
<td>SOM</td>
<td>5x5</td>
<td>68.39 ± 3.04</td>
</tr>
<tr>
<td>SOM</td>
<td>7x7</td>
<td>70.89 ± 2.92</td>
</tr>
<tr>
<td>XY-fused</td>
<td>3x3</td>
<td>66.25 ± 2.59</td>
</tr>
<tr>
<td>XY-fused</td>
<td>5x5</td>
<td>71.43 ± 2.23</td>
</tr>
<tr>
<td>XY-fused</td>
<td>7x7</td>
<td>73.75 ± 2.53</td>
</tr>
</tbody>
</table>

4 Results

4.1 Mental Stress Scenario

For each subject and each condition we have derived the mean values of 20 frequency bands from the spectra of the CoP. These feature vectors serve as input vectors $x_{stim}$ for the SOM and the XY-fused Kohonen Network respectively. The condition (stress or control) serves as class information for the Ymap of the XY-fused Kohonen Network. In order to obtain a reliable estimation of the discrimination accuracy, we have performed leave-one-person-out cross validations for both SOMs and XY-fused Kohonen Networks with varying node grid sizes of 3x3, 5x5 and 7x7. For each grid size, the cross validation was repeated 10 times in order to obtain an estimate about the variance of the overall classification error. In Tab. 2 the resulting classification accuracies including the standard deviations are shown. It can be seen that the increase in discrimination performance with rising grid size is consistent for the SOM as well as for the XY-fused Kohonen Network. The highest accuracy of 73.75% in discriminating stress from cognitive load is achieved with a 7x7 XY-fused Kohonen Network.

4.2 Emotion Recognition Scenario

Using the 25 utterance features as input vector and the emotions as output we have investigated XY-fused Kohonen Networks. In Fig. 3 an exemplary mapping of a network which was trained with data instances from the voice recording of all subjects is shown. It can be observed that most nodes contain instances from several emotions which render an accurate discrimination of different emotions impossible. There might be several reasons, why the training of the XY-fused Kohonen Network did not result in more homogeneous nodes. On the one hand, speech features are not only depending on the pronounced emotion but also on the speaker. Therefore we trained person-dependent XY-fused Kohonen Networks. In Fig. 4 a XY-fused Kohonen Network is shown which was trained with data instances from only one subject. It can be seen, that data instances from different emotions were mapped on different nodes. The only exception is the lower left node where beside data instances from emotion 6, one instance of the neighboring emotion 8 was mapped on the same node. Another reason for the inhomogeneous nodes of the XY-fused Kohonen Network trained with all data could be similarities between certain emotions. For example, in Fig. 3 it can be observed that the instances of the emotions pride and happiness are often mapped to the same cluster. We therefore trained a XY-fused Kohonen Network only with two emotions which we expected to be very different regarding their emotional arousal: desperation and boredom. The resulting mapping is shown in Fig. 5. It can be observed that most of the data instances from the emotion boredom were mapped on the upper part of the network while the instances from emotion disgust were mostly mapped on the lower part.
Figure 3: XY-fused Kohonen Network trained with utterance features from the voice recording of all subjects. The emotion classes are as follows: Disgust(2), Happiness(3), Cold anger(6), Boredom(7), Pride(8), Desperation(14).

Figure 4: Left: subject which expresses her emotions by speaking standard sentences. Right: XY-fused Kohonen Network trained with data instances from the voice recording of the subject shown on the left. The emotion classes are as follows: Disgust(2), Happiness(3), Cold anger(6), Boredom(7), Pride(8), Desperation(14).

5 Conclusion

In two experimental scenarios we have shown how SOMs and XY-fused Kohonen Networks can be employed to detect affective information. In a first scenario we have investigated the discrimination of mental stress and cognitive load. Based on features derived from a pressure mat, we have trained SOMs and XY-fused Kohonen Networks with varying node grid sizes. In a leave-one-person-out cross validation an overall discrimination accuracy of 73.75% could be achieved with a XY-fused Kohonen Network. This provides evidence that a person-independent discrimination of stress from cognitive load is feasible in an office scenario when using only pressure data.
In a second scenario we have investigated emotion detection from voice data. We have derived a total of 25 utterance features from the voice data as input for the XY-fused Kohonen Networks. It could be observed that in the resulting network most nodes contain instances from several emotions which render an accurate discrimination of different emotions impossible. In a person-dependent training, instances from different emotions were mapped in most cases on separate nodes. Considering only two emotions which differ regarding their emotional arousal resulted in two clearly separated regions in the network. These findings illustrate on the one hand that a person-dependent discrimination of emotions from voice data might be feasible. On the other hand, a general model might be appropriate to discriminate emotions which highly differ in the level of arousal.

Acknowledgments

The mental stress scenario was supported by a financial grant from the “Stiftung für Forschung und Nachwuchsförderung” of University Zurich. The emotion recognition scenario was supported by the European project MONARCA in the 7th Framework Programme under contract Number 248545.

References


