School Children Dyslexia Analysis using Self Organizing Maps

D. Novák\textsuperscript{1}, P. Kordík\textsuperscript{2}, M. Macaš\textsuperscript{1}, M. Vyhňálek\textsuperscript{3}, R. Brzezny\textsuperscript{3}, L. Lhotská\textsuperscript{1}

\textsuperscript{1}Department of Cybernetics, Czech Technical University in Prague, Czech Republic
\textsuperscript{2}Department of Computers, Czech Technical University in Prague, Czech Republic
\textsuperscript{3}Department of Neurology, 2nd Medical Faculty, Charles University, Czech Republic

Abstract—The main goal of the study is an unsupervised classification of school children dyslexia. Eye movements of 49 female subjects were measured using videooculographic technique (VOG) during two non-reading and one reading tasks. A feature selection was performed obtaining data set consisting of 26 features. Next an inductive modelling technique was applied to data set resulting in extraction of six features which were used as the input to self-organizing map (SOM). Three clusters were finally formed by the SOM proving that the proposed methodology is suitable for automatic dyslexia analysis.

Index Terms—Dyslexia Analysis, Inductive Modelling Feature Extraction, Self-Organizing Maps

I. INTRODUCTION

DYSLEXIA (specific reading disability) is a common, cognitively and behaviorally heterogeneous developmental condition, characterized primarily by severe difficulty in the mastery of reading despite average intelligence and adequate education [1]. Its prevalence in Czech school population is estimated to 2-3\% [2], which is considerably less than in English speaking countries (5-17.5\%) [3]. The aetiology of dyslexia is not fully understood but strong genetic background influenced by environmental factors have been demonstrated. The children with dyslexia need special educational methods for acquiring reading skills, that is why the early diagnostic is very important [4]. An early dyslexia detection will give those people the chance of being treated by means of the most accurate and specialized intensive therapy.

The numerous studies have demonstrated the abnormalities in structure and function of dyslexic brains. This brain dysfunction manifests also by eye movements abnormalities [5]. As it has been suggested, by measuring eye movements we could potentially predict children in risk for future development of dyslexia [6]. Up to present, only the direct inspection of eye movement traces by a skilled worker remains the most reliable method of learning disabilities detection. On the other hand, manual evaluation misses the important quantitative information about the vestibular excitability. With the aid of computer power, now available, it is possible to develop automated methods for early dyslexia and diagnose it in its very first stadium of development.

We propose the computerized method which is based on processing of the videooculography (VOG) eye movement signals. High correlation between the neuronal activity and eye response, as well as linear dependence of eye movement on stimulus velocity, has been clearly documented by numerous studies [7], [8]. Moreover, with the development of video and image analysis technology, various methods of automatic extraction of the eye position from images of the eye have been developed. Tracking the relative movements of these images gives an eye position signal. More commonly a video image is combined with computer software to calculate the position of the pupil and its centre. This allows vertical and horizontal eye movements to be measured. Regarding the methodology proposed, first we calculated a set of features from both horizontal and vertical eye movements signals. Next we applied a feature extraction method which is based on inductive modelling techniques. Having the most significant features at hand, a self organizing map (SOM) was applied for further unsupervised classification into three classes: patients with/without dyslexia and patients suffering from reading dysfunction.

II. METHODOLOGY

A. Data Acquisition

The eye movements of 49 female subjects were recorded using iView 3.0 videooculography system at Department of Neurology, 2nd Medical Faculty, Charles University, Czech Republic. The frequency rate of the camera used was 100Hz.

22 subjects were healthy, 18 subjects suffered from reading dysfunction and 9 subjects were dyslectic children. The average group age was 11 years, the age variance was 0.5 years. The subjects were stimulated by two non-verbal and one verbal stimuli. The non-verbal stimuli consist of the two images with different graphical patterns. The first image was the grid of blue dots (stimulus number 1), the subject was asked to inspect the image dot by dot. The second image was composed of the grid of yellow smileys, the subject was asked to count all smiling smileys (stimulus number 2). The next verbal stimulus was Czech text to read (stimulus number 3).

B. Feature Selection

Before feature selection itself, the data set was preprocessed as follows. First, the calibration of mapping on visual stimuli was carried out using projective geometric transformation. After calibration, the noise was filtered out of the signals by means of convolution filter with gaussian kernel. Finally, the blinking artefacts were removed. Then, the data set of 26 features were created from the VOG preprocessed horizontal and vertical signals over all three stimuli.

This work has been supported by the research program MSM 210000012 ‘Transdisciplinary Biomedical Engineering Research’ sponsored by the Ministry of Education, Youth and Sports of the Czech Republic.

2004 PROCEEDINGS OF THE 26TH ANNUAL EMBS INTERNATIONAL CONFERENCE, SEPTEMBER 3-5, SAN FRANCISCO, USA
presented. Sequences of fixations and saccades (rapid eye movements between fixations) define scanpaths, providing a record of visual attention on a display. By applying basic statistical operators on these two important characteristic (fixations and saccades) we obtained the feature data set as it can be seen in Table I. The definition of a fixation is a set of eye locations within a defined area for a selected amount of time. Eyes were allowed to drift during fixation as long as the amount of drift was less than this area per 100 ms.

### TABLE I

**Feature data set.**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS3</td>
<td>Reading speed of stimulus 3</td>
</tr>
<tr>
<td>RR3</td>
<td>Regression rate in stimulus 3</td>
</tr>
<tr>
<td>EE1,WE1,WW1,EW1</td>
<td>Direction of movement in stimulus 1</td>
</tr>
<tr>
<td>EE2,WE2,WW2,EW2</td>
<td>Direction of movement in stimulus 2</td>
</tr>
<tr>
<td>EE3,WE3,WW3,EW3</td>
<td>Direction of movement in stimulus 3</td>
</tr>
<tr>
<td>SA1,SA2,SA3</td>
<td>Average of ratio SA/SD in stimuli 1,2,3</td>
</tr>
<tr>
<td>SV1,SV2,SV3</td>
<td>Standard deviation of ratio SA/SD in stimuli 1,2,3</td>
</tr>
<tr>
<td>FA1,FA2,FA3</td>
<td>Average of fixation duration in stimuli 1,2,3</td>
</tr>
<tr>
<td>PV1,PV2,PV3</td>
<td>Standard deviation of fixation duration in stimuli 1,2,3</td>
</tr>
</tbody>
</table>

The first feature RS3 is a reading speed of the stimuli 3 (in milliseconds). Naturally, the dyslexia subjects should have this parameter bigger than the healthy ones. The second feature RR3 is the regression ratio defined in [5]. It is a sum of regression movement amplitudes divided by amplitude of accumulated amplitude of saccades during reading task. Next 12 features (EEi,WEi,WWi,EWi-in percentage) are the ratios of directions of eye movement from one fixation to the next. The proportion of each categorised direction and the transition matrix of each two successive categorised directions are then calculated. The categorised direction WE, for example, means that the eye entered from the previous fixation area to the actual one from the East direction and then the eye exited from the actual fixation area to the next one in the West direction. These measures were proposed as objective indicators of the strategies employed by an observer engaged in a visual search task [9]. Next 6 features SAIi and SVi are average and standard deviation of ratio \( \text{saccades amplitude/saccades duration} \) (SA/SD) in stimuli 1,2,3. For normal subjects the relationship between saccade amplitude and duration is fairly linear [10]. Finally, the last 6 features are average and standard deviation of fixation duration (in milliseconds) in stimuli 1,2,3. For example, according to [11], the average fixation duration of dyslectic German patient is 367 ms (standard deviation 132 ms), the healthy patient has the average fixation duration lower, 197 ms (standard deviation 34 ms).

### C. Feature Extraction

When modelling a real-world system, it is necessary to preselect a set of variables from the available information that may have impact on the behaviour of the system. By recording these variables in particular cases, a data set suitable for modelling can be produced.

The goal of variables selection is to avoid selecting too many or too few variables than necessary. In practical applications, it is impossible to obtain complete set of relevant variables. Siedlecki and Sklansky [12] used genetic algorithms for variable selection by encoding the initial set of \( n \) variables to a chromosome, where 1 and 0 represents presence and absence respectively of variables in the final subset. They used classification accuracy, as the fitness function and obtained good neural network results. These methods are usually used as a data preprocessing tool for further system modelling and classification by means of neural networks.

![Fig. 1. The example of the GAME network and the niching genetics algorithm evolving best performing units for the first layer of the network.](image)

We use an inductive approach combined with niching genetics algorithm instead of traditional neural networks. Whereas neural networks have a predefined structure, inductive models grow from the data set during the learning process. The construction process is highly efficient, it starts from the minimal form and the model grows according to system complexity. The resulting feed-forward hybrid network representing the model (see Figure 1) is generated by the Group of Adaptive Models Evolution (GAME) method [13].

### D. Self-Organizing Maps

The SOM is essentially a combined vector quantization and vector projection algorithm. It consists of neurons organized on a regular low-dimensional grid. Each neuron is represented by a \( d \)-dimensional weight vector, where \( d \) is equal to the dimension of the input vectors. The neurons are connected to adjacent neurons by a neighborhood relation, which dictates the topology, or structure, of the map. Typically the neurons are positioned on a 2-dimensional plane in a regular rectangular or hexagonal lattice.

The SOM is trained iteratively. In each training step, one sample vector \( x \) from the input data set is chosen randomly and the distance between it and all the weight vectors of the SOM is calculated using a distance measure. The neuron whose weight vector is closest to the input vector \( x \) is called...
the Best-Matching Unit (BMU) [14]. After finding the BMU, the weight vectors of the SOM are updated so that the BMU moved closer to the input vector in the input space. The training is usually performed in two phases. In the first phase, relatively large initial learning rate and neighborhood radius are used. In the second phase both learning rate and neighborhood radius are small right from the beginning. This procedure corresponds to first tuning the SOM approximately to the same space as the input data and then fine-tuning the map.

III. RESULTS AND DISCUSSION

The full data set (26 input variables, 3 output states corresponding to 3 classes, 49 vectors corresponding to 49 patients) was used to build 3 groups of nonlinear inductive models. First group models classified healthy patients, second group patients suffering from reading dysfunction and the third group patients with dyslexia.

In spite of applying the cross validation scheme, the inductive models overfitted data, loosing their modeling generalization abilities (see Figure 2).

Therefore we decided to use just linear units in inductive models. The classification accuracy dropped significantly, but we were able to visually study the influence of input variables to the output. Figure 3 shows that according to all linear inductive models, the growing reading speed increases the probability of the patient being healthy.

Our first experiments with SOM (full data set) did not reveal clusters separating healthy, dyslectic children and those suffering from reading dysfunction. We presumed that when just the most significant features for separating dyslectic children from others will be used, the result would be better. To extract such features we have used several methods implemented in the WEKA data mining software, but the best result was obtained by the method based on inductive modelling. We built the inductive model classifying the dyslectic children by the GAME method. During the construction stage of the model, we have accumulated number of units, connected to each input variable (for each layer of the model, after the niching genetic algorithm finished). These numbers in percent of overall connections for each feature are depicted in the Figure 4.

The most significant six features for dyslectic children classification were extracted (EE1,WE3,WW3,RS3,FV2,FV3). When the SOM was generated, some dyslectics were still mixed with healthy patients in the map. When we excluded the most significant feature RS3, the results were improved. As it can be seen in Figure 2, although the reading speed separates healthy/dyslectics well in average, it mixes those few healthy slow readers with dyslectic children. The experiment with the best visual cluster separation results, using extracted attributes (EE1,WE3,WW3,FV2,FV3) is described below.

Considering SOM experimental setup and implementation, the SOM Matlab toolbox was used. The SOM structure was two-dimensional hexagonal grid and the neighborhood function was Gaussian. The number of neurons were 42 forming $6 \times 7$ grid. A sequential algorithm was used to train the SOM with the initial learning rate 0.1, initial radius 5 and the number of training epochs was 8000. The random initialization was applied.

The most commonly used visualization technique to detect clusters from SOMs is distance u-matrix: a matrix of distance between neighboring map units. The units themselves (the hexagons) are coloured according to the median of the surrounding edges. Big values of the u-matrix correspond to a great distance between weight vectors of the map units, while small values mean that the map units are close to each other in the input space. Since big values are represented by dark colours, big gaps in the input space can be seen as dark borders between map units, while uniform areas can be seen
as light areas.

In Figure 5 the patients are assigned to each neuron using both labels in from of class-d-number and using color decoding. It can be seen that the SOM divides the input space to two areas where dyslexic patients are concentrated, mainly the cluster with patients numbered 60, 62, 63, 65, 67. Furthermore, it is not surprising that healthy subjects are mixed with subjects suffering from reading dysfunctions. Sometimes it is very difficult to separate clearly healthy subjects from subjects with reading difficulty.

**Fig. 5.** SOM clustering results. The red color (medium gray scale) is assigned to healthy subjects (label: 2-d-xx), green color (low gray) corresponds to patient with reading dysfunction (label: 3-d-xx) and finally the blue color (high gray) depicts patient with dyslexia (label: 4-d-xx).

Another technique is to visualize a feature planes. Each feature plane visualizes the spread of values of that feature. Coupled with clustering information, the feature planes show the values of the variables in each cluster as Figure 5 shows. It can be seen, for example, that high values of WW3 are typical for the cluster 'patient with reading dysfunctions' on the bottom of map (numbers: 46, 50, 55, 56, 76). Or, the low values of WE3 are significant for the dyslexia patient, e.g. left-upper part.

**Fig. 6.** Features significance using feature planes.

**IV. Conclusion**

We proposed an automatique technique for dyslexia analysis of school children. We showed that SOM is able to organize the data into three clusters which corresponds to subjects with/without dyslexia and with reading difficulties. However to get more consistent clustering we need to obtain more data, particularly more dyslectic patient should be measured.

During the testing procedure we used two non-reading tasks. It is important to point out that the two of the five most significant features were extracted from these non-verbal tests. Therefore in the future, there is possibility to analyse the dyslexia of pre-school children with the same methodology.

The inductive method GAME employed for the feature extraction also revealed hidden relationships in data. By visualizing the behavior of linear/nonlinear inductive models, we studied how the output variable (e.g. healthy) depends on any particular input variable (reading speed) in conditions given by constant values of other input variables.

Furthermore, we focussed on the selection of features with physiologic meaning only. We propose to use standard methods for feature selection (frequency analysis, wavelet transform, autoregressive methods, etc.) as a feature direction.

**Acknowledgment**

We are pleased to give our acknowledgment to Jakub Šnepok for creating the eye movement toolbox which was used for feature selection. Thanks also to Pau Micó for helpful hints on dyslexia problem.

**References**


