Using Recurrent Neural Networks for Automatic Chromosome Classification*

César Martínez, Alfons Juan, and Francisco Casacuberta

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, 46020 Valencia (Spain)
{cmargal,ajuan,fcn}@iti.upv.es

Abstract. Partial recurrent connectionist models can be used for classification of objects of variable length. In this work, an Elman network has been used for chromosome classification. Experiments were carried out using the *Copenhagen* data set. Local features over normal slides to the axis of the chromosomes were calculated, which produced a type of time-varying input pattern. Results showed an overall error rate of 5.7%, which is a good perfomance in a task which does not take into account cell context (isolated chromosome classification).

1 Introduction

The genetic constitution of individuals at cell level is the focus of cytogenetics, where genetic material can be viewed as a number of distinct bodies –the *chromosomes*–. In computer-aided imaging systems, which are now widely used in cytogenetic laboratories, the automatic chromosome classification is an essential component of such systems, since it helps to reduce the tedium and labour-intensiveness of traditional methods of chromosome analysis.

In a normal, nucleated human cell, there are 46 chromosomes represented in the clinical routine by a structure called *the karyotype*, which shows the complete set of chromosomes organized into 22 classes (each of which consists of a matching pair of two *homologous* chromosomes) and two sex chromosomes, XX in females or XY in males.

Producing a karyotype of a cell is of practical importance since it greatly facilitates the detection of abnormalities in the chromosome structure. Suitable chromosome staining techniques produce dark bands which are perpendicular to the longitudinal axis and are characteristic of the biological class thereby allowing its recognition. Automatic classification is based on this band pattern and other kinds of information (e.g., width) along the chromosomes. This work presents a pattern recognition application to the classification of microscopic greyscale images of chromosomes by connectionist systems, the Recurrent Neural Networks (RNN).

* Work supported by the Spanish "Ministerio de Ciencia y Tecnología" under grant TIC2000-1703-CO3-01.

J.R. Dorronsoro (Ed.): ICANN 2002, LNCS 2415, pp. 565–570, 2002.

2 Background

2.1 Introduction to the Pattern Recognition Problem

In the early days, chromosomes were stained uniformly and as a result they could be distinguished only on the basis of size and shape, the so-called *Denver groups*. However, nowadays most routine karyotyping is carried out on Giemsa-stained chromosomes. The chromosomes appear as dark images on a light background and have a characteristic pattern of light and dark bands which is unique to each class, and which is referred to as *G-banding*. Fig. 1 shows the karyogram of a normal metaphase cell, where the characteristic band pattern can be observed. It serves as a basis for the image processing and pattern recognition algorithms.

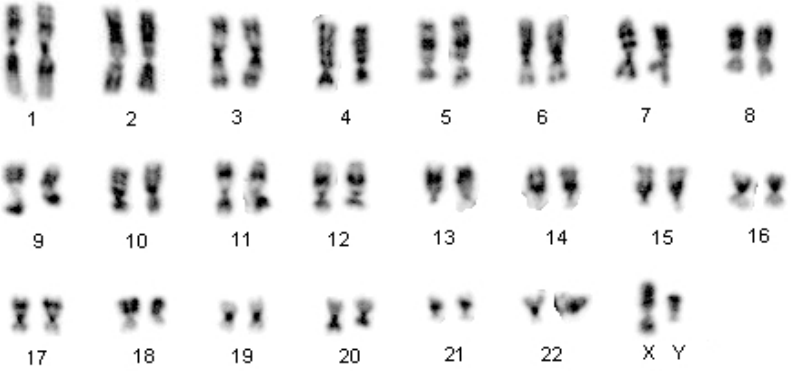

Fig. 1. Normal metaphase cell karyogram (images extracted from the Copenhagen data set).

2.2 Parameter Estimation

In 1989, Piper and Granum established a set of 30 features that were implemented in the chromosome analysis system MRC Edinburgh [1]. The features were classified according to how much a priori information is needed to measure them. Four feature *levels* were distinguished:

1. Direct measures on the image: area, density and convex hull perimeter.
2. Requirement of the axis: length, profiles of density, gradient and shape.
3. Requirement of axis plus polarity: asymetrical weight features.
4. Requirement of axis, polarity and centromere location: centromeric indices.

In this work, the parameters obtained from the images require the calculation of the longitudinal axis. The input patterns are built from measures over normal slices at unitary distance over the axis (details of the preprocessing methods are given in Section 3).

2.3 Recurrent Neural Networks

In many applications, time is inextricably bound up with many behaviors (such as language) and object representations. The question of how to represent time in connectionist models is very important, since it might seem to arise as a special problem which is unique to parallel processing models. This is because as the parallel nature of computation appears to be at odds with the serial nature of temporal events.

This work deals with the use of recurrent links in order to provide networks with a dynamic memory. In this approach, hidden unit patterns are fed back to themselves, acting as the context of prior internal states. Among the different networks proposed in the literature, the Elman network (EN) presents some advantages because of its internal time representation: the hidden units have the task of mapping both external input and the previous internal state (by means of the context units) to some desired output. This develops internal representations which are useful encodings of the temporal properties of the sequential input [7].

Fig. 2 shows an EN, where trainable connections are represented with dotted lines. Connections from the output of hidden units to the context layer are usually fixed to 1.0, and activations of hidden units are copied on a one-to-one basis. Basically, the training method involves activating the hidden units by means of input sequence segments plus prior activation of context units. Then, the output produced is compared with the teaching ouput and the backpropagation of the error algorithm is used to incrementally adjust connection strengths, where recurrent connections are not subject to adjustment [7].

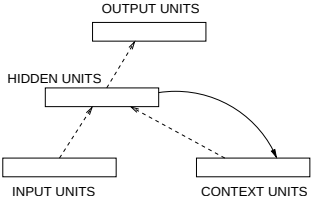

Fig. 2. A Recurrent Neural Network [7].

3 Materials and Methods

3.1 Corpus and Feature Extraction

The corpus used in the experiments was the *Cpa*, a corrected version of the large Copenhagen image data set *Cpr*. It consists of 2,804 karyotyped metaphase cells, 1,344 of which are female and 1,460 are male [4].

As a first step in the preprocessing stage, the histograms of the images are normalized and the holes are filled by means of contour chains to avoid problems in the axis calculation. Afterwards, mathematic morphology algorithms are applied (dilation-erosion) to smooth the chromosome outlines, which prevents the algorithm from producing an axis with more than two terminals.

Two skeleton techniques, the González-Woods algorithm [5] and the Hilditch algorithm [6], are successively applied to find the longitudinal medial axis. Finally, the transverse lines (slices), which are perpendicular to the axis are obtained.

A set of *local* features over the slices was calculated. The parameter set consisted of 9 bidimensional gaussian filtering points along the slices, all having an equidistant location between the slice bounds. At each point, a 5x5 pixel filter with the following weights was applied: 16 in the center, 2 at a in 1-pixel distance and 1 at a 2-pixel distance. Fig. 3 shows a chromosome along with an example of an input pattern resulting from 3 gaussian filtering points per frame (left, center and right points).

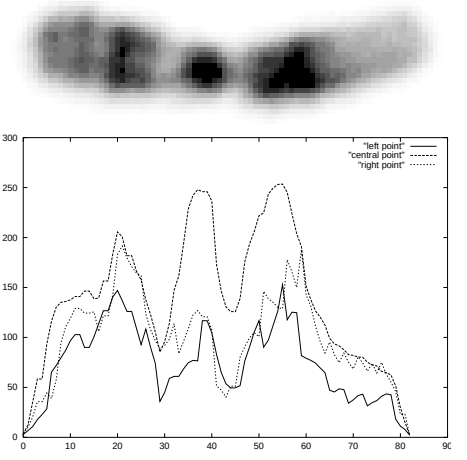

Fig. 3. Chromosome and 3-points gaussian profiles (chromosome image extracted from the Copenhagen data set).

3.2 Partial RNN

The SNNS toolkit provides the methods to implement different configurations of partial recurrent networks (Jordan and Elman networks) [2]. The EN implemented in this work had the following topology:

- 1 input layer of 45 units, in order to fit a 5-frame, moving window of 9 points each.
- 1 hidden layer, together with its context layer (variable number of units). The link weights from hidden units to context units were set to 1.0. There were no self-recurrent links for the context units.
- 1 output layer of 24 units, without context layer.

There is a special issue about testing the networks which should be highlighted. After the input frame presentation, the network produces a classification

in one of the 24 classes, so individual frames are classified based only on trained network knowledge. In order to test the EN perfomance, a special method was applied: a minimum number `MIN` of correct classified frames (in %) is established a priori. A chromosome is said to be correctly classified if more than `MIN` percent of frames are well-classified according to the desired output. For all the experiments, the value of `MIN` was fixed at 70%. This minimum was always surpassed in the slice voting scheme.

4 Experiments and Results

The same partition of Cpa corpus was used in all the experiments: the training set consisted of 33,000 chromosomes (It was not possible to use the full Cpa corpus size due to the high memory requirements of the feature set file). The validation set and the test set each consisted of 2,000 chromosomes. The number of units in the hidden layer was shifted from 50 to 250 (the same number of units was used in the context layer). The obtained error rate values are shown in Table 1.

Table 1. Test-set errors for different configurations of the hidden layer

Number of units	Classification error
50	22.8%
100	15.3%
150	9.5%
200	5.7%
250	6.7%

The best perfomance was achieved with 200 units in the hidden layer, which obtained a mean error rate for the 24 classes of 5.7%. The lowest individual error rate was 3.1% (for class 3), whereas the highest error rate was 40.0% (for class 24, which corresponds to sex chromosome Y).

5 Discussion and Conclusions

This work presents the application of partial recurrent neural networks (specifically Elman networks) to automatic chromosome classification. Local information from the length of the chromosome was extracted from the greyscale images. This information (gaussian filtering points sampled at equidistant distance over normal lines to the axis) constitute the local feature frame.

The best result achieved was a minimum error rate of 5.7% in isolated chromosome classification, using a test-set of 2,000 input patterns. This was achieved by means of a standard and well-known Neural Network technique. The use of

recurrent neural networks is supported given the fact that feedforward netowks, in the same conditions, achieved a minimum error rate of 12.88%.

This preliminary technique could be improved in further research studies. For example, a natural step for improving the error rate would be to restrict the system to doing classification by cell, exploiting the fact that a normal cell has only 46 chromosomes ordered in 24 classes. Ritter *et al* use this information and other improvements such as an enhanced profile extraction method to obtain an error rate of 0.61% using a Bayesian classifier [8]. The human error rate of an experienced cytogenetist lies between 0.1% (using good quality images) and 0.3% (clinical applications).

The higher individual error rate obtained for class 24 (sex chromosome Y, only present in normal male cells) was expected because of its lower a-priori probability: 1/92 compared to 3/92 for class 23 (sex chromosome X, present in both male and female cells) and 4/92 for classes 1 to 22.

The same partition of training-validation-test was used for all the experiments in this work due to high computational requirements. However, the application of a m-fold cross-validation method would be better to obtain statistically representative results. Experiments using the complete Cpa corpus should be done to improve the results.

Acknowledgments. The authors wish to thank Dr. Gunter Ritter (*Fakultät für Mathematik und Informatik, Universität Passau, Passau, Germany*) for kindly providing the Cpa chromosome-image database used in this work.

References

1. J. Piper: Some properties of six chromosome data sets. Medical Research Council Report, MRC Human Genetics Unit, Edinburgh (1990)
2. SNNS Stuttgart Neural Network Simulator v4.2, University of Stuttgart - University of Tübingen, Germany, 1998.
   `ftp.informatik.uni-tuebingen.de/pub/SNNS`
3. J. Piper: Variability and bias in experimentally measured classifier error rates. Pattern Recognition Letters **13** (1992) 685-692.
4. G. Ritter, G. Schreib: Using dominant points and variants for profile extraction from chromosomes. Patter Recognition **34** (2001) 923-938
5. R. Gonzalez, R. Woods: Digital image processing. Addison-Wesley Publishing Company (1992)
6. C. Hilditch: Linear skeletons from square cupboards. Machine Intelligence **19** (1969) 403-420
7. J. Elman: Finding Structure in Time. Center for Research in Language, University of California, San Diego (1988)