Interactive Structured Output Prediction: Application to Chromosome Classification

Jose Oncina[1] and Enrique Vidal[2]

[1] Dept. Lenguajes y Sistemas Informáticos
Universidad de Alicante
oncina@dlsi.ua.es
[2] Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
evidal@iti.upv.es

Abstract. Interactive Pattern Recognition concepts and techniques are applied to problems with structured output; i.e., problems in which the result is not just a simple class label, but a suitable structure of labels. For illustration purposes (a simplification of) the problem of Human Karyotyping is considered. Results show that a) taking into account label dependencies in a karyogram significantly reduces the classical (non-interactive) chromosome label prediction error rate and b) they are further improved when interactive processing is adopted.

Keywords: interactive pattern recognition, machine learning, structured output prediction, chromosome classification, karyotype recognition

1   Introduction

Classification is one of the most traditional Pattern Recognition (PR) frameworks [2]. For a given input x, the set of possible output hypotheses is a finite (and typically small) set of class-labels, or just integers $\{1, \ldots, C\}$, where C is the number of classes. In this case, the *search* needed to solve the recognition problem amounts to a straightforward exhaustive exploration of the corresponding C posterior probability values, $\Pr(h \mid x)$; that is,

$$\hat{h} = \operatorname*{arg\,max}_{1 \leq h \leq C} \Pr(h \mid x) \tag{1}$$

While classification is in fact a useful framework within which many applications can be naturally placed, there are many other practical problems of increasing interest which need a less restrictive framework where hypotheses are not just labels, but some kind of *structured* information. This is the case, for example, of Automatic Speech or Handwritten Text Recognition (ASR, HTR), Machine Translation (MT), etc. In these cases, the inputs, x, are structured as *sequences* of feature vectors (ASR, HTR) or words (MT) and the outputs, h, are *sequences* of words or other adequate linguistic units. Many applications admit this kind of input and output sequential structuring, but there are also other

practical problems, many of them in the field of Image Processing and Computer Vision, which require more complex structures such as input and output *arrays* or *graphs* of vectors and labels, respectively.

Let $\mathcal{H}$ be the *structured hypotheses space*. Now (1) is written as:

$$\hat{h} = \underset{h \in \mathcal{H}}{\arg\max} \Pr(h \mid x) \tag{2}$$

Depending on the exact nature of $\mathcal{H}$, this optimization can become quite complex, but several adequate algorithmic solutions or approximations, such as *Viterbi search* [12,3], $A^{\star}$ [1,7], etc., have been developed over the last few decades.

In this paper we are interested in applying Interactive PR (IPR) [11] approaches to problems with structured output because it is in this kind of problems where the IPR framework is likly to be most fruitful.

To illustrate concepts, problems and approaches in this framework, we will consider here a simplification of a classical PR problem: the recognition of human karyotypes. While individual chromosome recognition [10,5] is a typical PR example of *classification*, the recognition of a whole karyotype [8,6] properly corresponds to the case of *structured input/output*, as will be discussed below.

A *karyotype* is the number and appearance of chromosomes in the nucleus of a eukaryote cell. Normal human karyotypes contain 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Normal karyotypes for females contain two X chromosomes, males have both an X and a Y chromosomes. Any variation from the standard karyotype may lead to developmental abnormalities. The chromosomes are depicted (by rearranging a microphotograph) in a standard format known as a *karyogram* or idiogram: in pairs, ordered by size and position of centromere for chromosomes of the same size[3]. Each chromosome is assigned a label from $\{$"1", ..., "22", "X", "Y"$\}$, according with its position in the karyogram [8].

In this work we consider the problem of karyotype recognition and we explore IPR approaches to increase the productivity with respect to a traditional, *non interactive* or *"offline"* PR approach.

In order to focus on the most relevant aspects of the problem we will not consider the real, full karyotype recognition problem, but a simpler setting in which only single chromosome images, rather than pairs, are considered and sex chromosomes, "X", "Y", are ignored. Then a *karyotype* is represented by a sequence or vector of chromosome images $\mathbf{x} = (x_i)_{i=1}^{22}$. Our task is to obtain the corresponding *karyogram*; i.e., a corresponding sequence or vector $\mathbf{h} = (h_i)_{i=1}^{22}$, where each h_i is the label or class of the chromosome image x_i, $i \in \{1, \ldots, 22\}$. For example, $h_4 = 7$ means that the chromosome image x_4 belongs to class 7 or has the label "7" in the karyogram.

---

[3] For the sake of simplicity, we ignore here the initial image segmentation task and assume that each of the 46 chromosomes in a normal unsorted karyotype is already represented as an individual image. Moreover, we do not take into account recent advances in karyotype analysis, such as fluorescent dye based spectral karyotyping [9], which allow obtaining coloured chromosome images and may significantly simplify the real human karyotyping problem.

From now on, we assume that a reliable PR system is available for classifying individual chromosome images. For each image x_i, $i \in \{1, \ldots, 22\}$, the system provides us with adequate approximations, $P(j \mid x_i)$, to the posterior probabilities $\Pr(j \mid x_i)$, $j \in \{\text{"1"}, \ldots, \text{"22"}\}$.

2   Non-interactive and Interactive frameworks

We explore three different frameworks: One non interactive or *"offline"* and two interactive called *"active"* and *"passive"*.

The names *active* and *passive* refer to who takes the supervision "initiative". In the active case, the system "actively" proposes items to supervise, while in the passive case, it just "passively" waits for the user to decide which items need supervision and/or correction.

Offline: The system proposes a vector of labels **h**. This vector is supervised by a user who corrects all the errors. User's effort is measured in two ways: a) the number of karyograms with at least one misclassified chromosome. In this case we are assuming that the same effort is needed to correct a single chromosome label as to correct several; b) the number of misclassified chromosomes. We assume that the effort is proportional to the number of label corrections needed.

Passive: The system proposes a karyogram hypothesis **h**. Then the user examines its labels $h_1, h_2, \ldots$ sequentially until the first error is found. After correcting this error, the system proposes a new karyogram consistent with all the previously checked and/or corrected elements. Note that this protocol can be equivalently formulated as follows: The system, sequentially for $i = 1, \ldots, 22$, proposes the candidate label h_i for the chromosome image x_i. At each step, the user corrects the possible label error. In this framework the obvious measure of effort is counting the number of corrections the user has to make. However, we will also report the number of karyograms that need at least one correction.

Active: The system sequentially proposes a pair (i, j) as an hypothesis that the chromosome x_i is of the class $h_i = j$ in the karyogram. Like in the previous case, the effort is measured as the number of times the user should correct the possible system hypothesis error.

3   Offline framework

In this case, classical, non-interactive processing is assumed. Different scenarios are considered, depending on which errors we want to minimize.

3.1   Offline Individual Chromosomes

This is perhaps the simplest setting in which individual chormosome images have to be classified without taking into account that they may belong to a karyotype. This is the setting we find in the majority of PR papers dealing with chromosome recognition (e.g., [10,5]).

In traditional PR [2], decision theory is used to minimize the cost of wrong hypotheses. A 0/1 cost or *loss* function corresponds to minimizing the number of wrong hypotheses. Under this minimal error loss, the best hypothesis is shown to be one which maximises the hypothesis posterior probability.

In this case, the individual chromosome error is minimised by maximizing the posterior probability for each chromosome image; that is, for all $\mathbf{x}$ and for each $i \in \{1, \ldots, 22\}$:

$$\hat{h}_i = \underset{j \in \{1,\ldots,22\}}{\arg\max} \ P(j \mid x_i) \tag{3}$$

3.2 Offline Karyotype Global

Here we aim to minimize complete-karyogram errors. According to decison theory, for each $\mathbf{x}$ we have to search for the most probable karyogram, $\hat{\mathbf{h}}$; that is:

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}}{\arg\max} \ \mathrm{Pr}(\mathbf{h} \mid \mathbf{x}) \tag{4}$$

Assuming independence beyond the impossibility of assigning two different labels to the same chromosome image, we can write:

$$\mathrm{Pr}(\mathbf{h} \mid \mathbf{x}) = \begin{cases} C \prod_{i=1}^{22} \mathrm{Pr}(h_i \mid x_i) & \text{if } \mathbf{h} \in \mathcal{H}' \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where C is a normalization constant and $\mathcal{H}' = \{\mathbf{h} \in \mathcal{H} : h_i \neq h_j \ \forall i \neq j\}$ is the set of valid hypothesis (those without repeated labels). This way (4) becomes:

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}'}{\arg\max} \prod_{i=1}^{22} P(h_i \mid x_i) \tag{6}$$

To approximately solve this difficult maximization problem, a greedy strategy is adopted. First we compute $(\hat{i}, \hat{j}) = \arg\max_{i,j} P(j \mid x_i)$ and we assign $h_{\hat{i}} = \hat{j}$. Then, we eliminate the chromosome $x_{\hat{i}}$ and label $\hat{j}$ from the $\arg\max$ searching set and repeat the process until all elements of $\mathbf{h}$ have been assigned.

3.3 Offline Karyotype Unconstrained

This setting is similar to the previous one in that each batch of 22 chromosome images, $\mathbf{x}$, is considered to be a complete karyotype. But here we aim to minimize the number of chromosome (rather than complete-karyogram) errors.

Let $\mathbf{h}$ be a proposed hypothesis and $\mathbf{h}^\star$ the "correct" hypothesis. The *loss* function in this case is not 0/1, but the total number of missclassified chromosomes in $\mathbf{h}$. This loss is given by the Hamming distance:

$$d(\mathbf{h}, \mathbf{h}^\star) = \sum_{1=1}^{22} [h_i \neq h^\star{}_i] \tag{7}$$

where $[\mathcal{P}]$ denotes the Iverson bracket, which is 1 if $\mathcal{P}$ is *true* and 0 otherwise. Then, the conditional risk [2] (i.e., the expected number of errors when a hypothesis $\mathbf{h}$ is proposed for a given $\mathbf{x}$) is:

$$R(\mathbf{h} \mid \mathbf{x}) = \sum_{\mathbf{h}' \in \mathcal{H}} d(\mathbf{h}, \mathbf{h}') \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{8}$$

and the hypothesis that minimises this risk is:

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}}{\arg\min} \sum_{\mathbf{h}' \in \mathcal{H}} d(\mathbf{h}, \mathbf{h}') \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{9}$$

$$= \underset{\mathbf{h} \in \mathcal{H}}{\arg\min} \sum_{i=1}^{22} \sum_{\mathbf{h}' \in \mathcal{H}} [h_i \neq h_i'] \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{10}$$

$$= \underset{\mathbf{h} \in \mathcal{H}}{\arg\max} \sum_{i=1}^{22} \sum_{\mathbf{h}' \in \mathcal{H}} [h_i = h_i'] \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{11}$$

And now, since the i-summation terms are independent, the maximisation can be split into 22 maximization problems, one for each h_i.

$$\hat{h}_i = \underset{j \in \{1, \dots, 22\}}{\arg\max} \sum_{\substack{\mathbf{h} \in \mathcal{H} \\ h_i = j}} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{12}$$

$$= \underset{j \in \{1, \dots, 22\}}{\arg\max} \sum_{\substack{\mathbf{h} \in \mathcal{H}' \\ h_i = j}} \prod_{k=1}^{22} P(h_k \mid x_k) \tag{13}$$

$$= \underset{j \in \{1, \dots, 22\}}{\arg\max} \, P(j \mid x_i) \sum_{\substack{\mathbf{h} \in \mathcal{H}' \\ h_i = j}} \prod_{\substack{k=1 \\ k \neq i}}^{22} P(h_k \mid x_k) \tag{14}$$

Finally, it is interesting to see that, if we assume in (14) that the individual chromosome probabilities are reasonably well approximated, the summation would not vary enough to dominate the big variations of $P(i \mid x_j)$ and, therefore,

$$\hat{h}_i \approx \underset{j \in \{1, \dots, 22\}}{\arg\max} \, P(j \mid x_i) \tag{15}$$

which is identical to the classical solution to the *Offline Individual Chromosome* setting. Note that, as in that setting, here we are not restricting $\hat{\mathbf{h}}$ to be a valid hypothesis. That is, in the optimal $\hat{\mathbf{h}}$ we may have $\hat{h}_i = \hat{h}_j, i \neq j$.

We can enforce finding only valid hypothesis through a simple heuristic: at each step, select the chromosome label that maximises $P(j \mid x_i)$, provided j was not used in a previous step. It may be argued that introducing this restriction will lead to more accurate predictions. However, with the approximation (15), this heuristic exactly leads to the greedy solution to the *Offline Karyotype Global* problem discussed at the end of section 3.2.

4 Interactive Passive framework

In this framework two approaches have been considered: *Karyotype* and *Karyotype Unconstrained*. In both cases, it is assumed that the karyogram elements are explored in a *left–to–right* sequential order. In what follows, suppose we are at the ith interaction step and let $\mathbf{h'}$ denote the hypothesis provided by the system in the previous step, $i-1$. Given the left–to–right exploration, all the elements $h_1'^{i-1}$ of $\mathbf{h'}$ are known to be correct.

4.1 Interactive Passive Left–to–Right Karyotype Global

In this strategy we look for a hypothesis, compatible with the known correct information in $h_1'^{i-1}$, which minimises the expected number of whole karyogram errors. That is:

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}}{\arg\max}\, \Pr(\mathbf{h} \mid \mathbf{x}, h_1'^{i-1}) \tag{16}$$

$$= \underset{\substack{\mathbf{h} \in \mathcal{H} \\ h_1^{i-1} = h_1'^{i-1}}}{\arg\max}\, \Pr(\mathbf{h} \mid \mathbf{x}) \tag{17}$$

$$= \underset{\substack{\mathbf{h} \in \mathcal{H}' \\ h_1^{i-1} = h_1'^{i-1}}}{\arg\max}\, \prod_{k=i}^{22} P(h_k \mid x_k) \tag{18}$$

As in *Offline Karyotype Global*, a greedy approach is used for this maximisation. First all the labels known from the previous step are assigned; i.e., $h_1^{i-1} = h_1'^{i-1}$. Next we obtain $(\hat{k}, \hat{j}) = \arg\max_{i \le k \le 22, j \notin h_1'^{i-1}} P(j \mid x_k)$ and assign $\hat{h}_{\hat{k}} = \hat{j}$. Then, the chromosome $x_{\hat{k}}$ and the label $\hat{j}$ are removed from the searching set and the process is repeated until all elements of $\hat{\mathbf{h}}$ have been assigned.

4.2 Interactive Passive Left–to–Right Karyotype Unconstrained

In this case, at each step i^{th} we just look for the most probable label for the ith chromosome image, assuming all the labels assigned in previous steps are correct. Clearly, in this way we do not explicitly care about possible label repetitions for the labels to be assigned in further steps and this is why this strategy is called *"unconstrained"*. However, since the single label to be assigned at each step is restricted to be different from those assigned in previous steps, the final result obtained at the end of the process is guaranteed to be valid karyogram.

Formally, we look for the most probable label h_i for the chromosome image x_i, given that all the labels $h_1^{\prime i-1}$ of $\mathbf{h}'$ are correct. That is:

$$\hat{h}_i = \underset{j \in \{1,\dots,22\}}{\arg\max} \sum_{\substack{\mathbf{h} \in \mathcal{H} \\ h_i = j}} \Pr(\mathbf{h} \mid \mathbf{x}, h_1^{\prime i-1}) \tag{19}$$

$$= \underset{j \in \{1,\dots,22\}}{\arg\max} \sum_{\substack{\mathbf{h} \in \mathcal{H} \\ h_i = j \\ h_1^{i-1} = h_1^{\prime i-1}}} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{20}$$

$$= \underset{j \in \{1,\dots,22\}}{\arg\max} \; P(j \mid x_i) \sum_{\substack{\mathbf{h} \in \mathcal{H}' \\ h_i = j \\ h_1^{i-1} = h_1^{\prime i-1}}} \prod_{k=i+1}^{22} P(h_k \mid x_k) \tag{21}$$

As in the *Offline Karyotype Unconstrained* case, if we assume that the summation is going to change less than $P(j \mid x_i)$. Then,

$$\hat{h}_i \approx \underset{\substack{j \in \{1,\dots,22\} \\ j \notin h_1^{\prime i-1}}}{\arg\max} \; P(j \mid x_i) \tag{22}$$

4.3 Interactive Active framework

In this framework, at the step i^{th}, the system chooses which chromosome and class label has to be supervised. In the previous karyogram, $\mathbf{h}'$, we write $h'_k = 0$ if an only if we don't know whether the kth label in $\mathbf{h}'$ is correct. Let $c(\mathbf{h}') = \{j : j = h_k \neq 0, \; 1 \leq k \leq 22\}$ be the set of correct labels in $\mathbf{h}'$. An optimal chromosome-label pair to be supervised is:

$$(\hat{k}, \hat{j}) = \underset{\substack{k:h_k=0 \\ j \notin c(\mathbf{h}')}}{\arg\max} \sum_{\substack{\mathbf{h} \in \mathcal{H} \\ h_k = j}} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{23}$$

$$= \underset{\substack{k:h_k=0 \\ j \notin c(\mathbf{h}')}}{\arg\max} P(j \mid x_k) \sum_{\substack{\mathbf{h} \in \mathcal{H}' \\ h_k = j}} \prod_{\substack{l=1 \\ l \neq k}}^{22} P(h_l \mid x_l) \tag{24}$$

As in previous cases, if the variation is dominated by $\Pr(j \mid x_k)$:

$$(\hat{k}, \hat{j}) \approx \underset{\substack{(k,j) \\ h'_i \neq 0, j \notin c(\mathbf{h}')}}{\arg\max} \; P(j \mid x_k) \tag{25}$$

5 Experiments

The experiments presented in this work have been carried out using the so-called *"Copenaghen Chromosomes Data Set"*. The raw data, preprocessing and representation are described in detail in [4,10]. Chromosome images are finally represented as variable-length strings of symbols, each of which represents the

variation of the image grey-level along the chromosome median axis. The centromere position was marked using a special symbol at the corresponding string position [10]. In total, 200 karyotypes and 4,400 chromosome samples are available in this data set.

These samples were split into two balanced blocks of 100 karyotipes (2,200 chromosome samples) and every experiment entailed two runs following a two-blocks Cross-Validation scheme. The classification error-rates reported below are the average result of these two runs.

The probabilities needed to apply the methods described in the previous sections where obtained with the so-called ECGI approach [10]. Models used in this approach can be seen as a kind of Hidden Markov Models where the topology is automatically derived from the training strings. For each chromosome class $j \in \{1, \ldots, 22\}$, a model was trained using the training strings of this class. Then, for each test string x, its corresponding 22 class-likelihoods $P(x \mid j)$ were computed by parsing x through the 22 trained models. The posterior probabilities $P(j \mid x), j \in \{1, \ldots, 22\}$, were obtained by normalizing the likelihoods assuming uniform priors for the 22 possible classes.

Using these probabilities, the error rate for individual chromosome classification was 5.7% [10]. This is the baseline for the experiments here presented.

The following methods have been tested: *"Offline Individual Chromosomes"* (OIC, eq. 3), *"Offline Karyotype Global"* (OKG, eq. 6), *"Interactive Passive Left-to-Right Karyotype Unconstrained"* (IPU, eq. 22), *"Interactive Passive Left-to-Right Karyotype Global"* (IPG, eq. 18) and *"Interactive Active"* (IAC, eq. 25). The *"Offline Karyotype Unconstrained"* framework has not benn tested because, as noted in Section 3.3, approximations and greedy solutions make solutions for this framework identical to those of *"Offline Individual Chromosomes"* or *"Offline Karyotype Global"*.

It is worth noting that all these methods, except IPU (eq. 22), are insensible to the order in which the chromosomes appear in **x**. Nevertheless, each experiment has been carried out twice, one with the original order of the chromosomes in the data set and another with the reverse of this order. Results are averaged for these two runs.

6 Results

Empirical results are shown in Table 1. As expected the *Global* methods lead to the best karyotype-level results (and also at the chromosome level). On the other hand, interactive processing clearly requires far fewer label correction: 44% fewer for both PKU relative to OIC and PKG relative to OKG. Finally, the IAC approach achieves the overall best results.

7 Discussion and Conclusions

This work shows how to apply interactive Pattern Recognition concepts and techniques to problems with structured output. For illustration purposes these

Table 1. Karyotype and chormosome error corrections needed (in %).

Method	Equation	Karyotype	Chromosome
Offline Indivual Chromosome (OIC)	(3)	56	5.7
Offline Karyotype Global (OKC)	(6)	27	3.7
Pasive Karyotype Unconstrained (PKU)	(22)	40	3.2
Pasive Karyotype Global (PKG)	(18)	27	2.1
Active (IAC)	(25)	27	1.9

techniques are applied to (a simplification of) the problem of Human Karyotyping. Results show that a) taking into account label dependencies in a karyogram significantly reduces the classical (noninteractive) chromosome label prediction erros and b) performance is further improved when interactive processing is adopted. These results have been obtained using both search and probability computation approximations. Further improvements are expected by improving the accuracy of these computations.

References

1. Dechter, R., Pearl, J.: Generalized best-first search strategies and the optimality of A*. J. ACM 32, 505–536 (July 1985), http://doi.acm.org/10.1145/3828.3830
2. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley (1973)
3. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1998)
4. Kao, J.h., Chuang, J.h., Wang, T.: Chromosome classification based on the band profile similarity along approximate medial axis. Pattern Rec. 41, 77–89 (2008)
5. Martínez, C., , García, H., Juan, A.: Chromosome classification using continuous hidden markov models. In: Pattern Recognition and Image Analysis, pp. 494–501. LNCS, Springer Verlag, (2003)
6. Martínez, C., Juan, A., Casacuberta, F.: Iterative contextual recurrent classification of chromosomes. Neural Processing Letters 26(3), 159–175 (2007)
7. Pearl, J.: Heuristics - Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley Reading, Mass (1985)
8. Ritter, G., Gallegos, M., Gaggermeier, K.: Automatic context-sensitive karyotyping of human chromosomes based on elliptically symmetric statistical distributions. Pattern Recognition 28(6), 823–831 (1995)
9. Schröck, E., du Manoir, S., Veldman, T., B. Schoell, J.W., Ferguson-Smith, M.A., Y. Ning, D.H.L., Bar-Am, I., Soenksen, D., Garini, Y., Ried, T.: Multicolor spectral karyotyping of human chromosomes. Science 273(5274), 494–497 (1996)
10. Vidal, E., Castro, M.: Classification of Banded Chromosomes using Error-Correcting Grammatical Inference (ECGI) and Multilayer Perceptron (MLP). In: A.Sanfeliu, J.Villanueva, J.Vitriá (eds.) In VII National Symposium on Pattern Recognition and Image Analysis. pp. 31–36. Barcelona (1997)
11. Vidal, E., Rodríguez, L., Casacuberta, F., García-Varea, I.: Interactive pattern recognition. In: Proceedings of the 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, LNCS, vol. 4892, pp. 60–71. Springer Verlag, Brno, Czech Republic (June 2007)
12. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Transactions on Information Theory 13, 260–269 (1967)