ON KNOWLEDGE-BASED IMPROVEMENT OF
BIOMEDICAL PATTERN RECOGNITION
- A CASE STUDY

Q. Wu, P. Suetens, A. Oosterlinck
ESAT-MI2, Dept. Electrical Engineering,
University of Leuven, Kardinaal Mercierlaan 94,
B-3030 Heverlee, Belgium
(Tel. 32-16-286211-ext.2250)

Case Study: Knowledge-Based Pattern Recognition

ABSTRACT: Most biomedical pattern recognition (BPR) systems use the classical statistical pattern recognition strategy in which a feature hyperspace is constructed for the problem followed by a statistical discriminant analysis. In this paper however it is shown that there are several essential drawbacks with this conventional approach. Chromosome analysis for example, which has long been an outstanding BPR problem, suffers from the practical limitations of such an approach. In a recent pilot study on applying knowledge-based system techniques to chromosome classification, we propose a novel scheme to overcome some of these limitations, aiming to come up with a strategy for knowledge-based improvement of biomedical pattern recognition in general. By contrast to the single-path decision theoretical approach used in BPR systems, a multistage recursive hypothesize-and-verify paradigm is adopted in our scheme. The experience invested in classical statistical pattern recognition is retained and used to generate hypothetical classifications, and the domain expert knowledge can be incorporated to form a 'constraint network', which is implemented in terms of rule-based updating of beliefs and verification of hypothetical classifications. Preliminary tests of a chromosome analysis system using this scheme have been carried out on metaphase image data. The results presented indicate that an encouraging improvement of classification performance can be achieved, using the proposed scheme over the conventional BPR approach.

STATUS: Research and Field Test

DOMAIN: Automated Cytogenetics

LANGUAGE: C and OPS5

EFFORT: About One Year and a Half
Introduction

Current research in artificial intelligence has indicated that the key to the creation of intelligent systems is the incorporation of large amounts of task-specific knowledge. Knowledge-Based System (KBS) or Expert System (ES) techniques as spin-offs of applicational AI research come to play an increasingly important part in building computers that exploit the specialized knowledge of experts to achieve high performance in the specific problem domain. Typically the field of biomedical information processing has been regarded as an appropriate testbed for investigating KBS or ES techniques, as it is generally believed that decision making in biomedical expert reasoning usually involves the use of heuristics from domain knowledge sources [1-3]. Most biomedical pattern recognition (BPR) systems to date however, use the classical statistical pattern recognition strategy in which a feature hyperspace is constructed for the problem followed by a statistical discriminant analysis procedure. In the following sections of this paper it is shown that there are several essential drawbacks with this conventional approach. Chromosome analysis for example, which has long been an outstanding BPR problem, suffers from the practical limitations of such an approach. These limitations often lead to highly erroneous classification results subject to time-consuming interactive corrections in existing systems [4-9]. In order to overcome such limitations, recently a pilot study on applying KBS techniques to chromosome classification has been carried out, in which the following major issues were investigated:

- What kind of domain knowledge, and how this knowledge can be represented and used to achieve improvement of chromosome classification performance.
- Which problem solving paradigm, and which inference mechanism are to be used in the system.
- How to retain and utilize the experience invested in classical statistical pattern recognition, so that the system could be built on top of the current performance level instead of from scratch.
- For a practical implementation of the system, given the fact that knowledge acquisition and validation goes on in an evolutionary way, how to implement the augmented knowledge-based component as 'plug-in' modules for easy incremental modification.

A new scheme was developed as a result of this case study, and is described in detail in the remaining of the paper. The system is implemented and tested on metaphase imagery in our preliminary experiments, the results of which are also presented.

The conventional BPR approach

In general, most BPR systems use the classical statistical pattern recognition approach [10-13] that can be characterized by a block diagram (Figure 1) illustrating the functional and control flow of their processes.

The sensor data correspond to biomedical signals in different formalities and dimensionalities, e.g., ECG (Electrocardiogram), EEG (Electroencephalogram), microscopic or X-ray images, etc., which are subject to preprocessing steps ranging from digitization, noise cleaning to segmentation. In a chromosome analysis system, the input data are the metaphase images digitized from microscopic view of metaphase cells (Figure 2).

However in the following context we restrict our discussion from the derived patterns onward to classification stage and concern only points raised from classification strategies, as this being the scope in which the use of KBS techniques has been attempted.

A pattern is usually defined as an individual basic object unit (segmented from the sensor data) under classification in the domain of discourse, say, an individual chromosome represented by a matrix of quantized samples (pixel) of gray-level intensities. After some feature reduction or extraction processes, it is further represented by an n-dimensional feature vector:

\[ F[f_1, ..., f_n] \]

where n stands for the number of features measured from each pattern, e.g., each chromosome in the image. In this way a n-dimensional feature hyperspace is constructed for further discriminant analysis. In statistical pattern recognition, to classify a pattern into one of m classes, an estimation of multiclass conditional probability density functions has to be made from the learning sample data:

\[ P(F|C_j), j = 1, ..., m \]

And the calculation of the likelihood functions results in:

\[ L_j = P(C_j|F) = \frac{P(F|C_j)P(C_j)}{P(F)} \]

with \( P(F) \) a normalizing factor. According to Bayesian decision rule, the optimal classification can be achieved by minimizing wrong classification (error), i.e., assigns a pattern to the class with the highest likelihood or a posteriori probability. In chromosome classification this means assigning a chromosome i into class j when

\[ L_{ij} \geq L_{ik}, k = 1, ..., m, k \neq j \]

However, the above approach exhibit several essential drawbacks as can be conceived of and substantiated by the following:
Problems arise in obtaining good estimation of the a priori probability and conditional probability density functions. The true distribution of the random variable $F$ (feature vector) is almost never known due to insufficient size of learning set of classified patterns. For the chromosome classification problem which exhibits serious biological variations, the (n-dimensional) multiclass conditional probability density functions of each object $P[F | C_j]$ can hardly be reliably estimated in reality from learning sample set for the formidable size requirement. Thus in practice either a statistical independence of each feature has to be assumed so that

$$P[F | C_j] = \prod_{i=1}^{n} P[f_i | C_j], i = 1, ..., n$$

and the estimation can be facilitated by learning on each feature separately, or an educated guess at the form of statistical models (e.g., normal distributions) of multi-variate probability density functions is made, by which we only have to estimate parameters describing the the functions. These treatments are mainly on account of simplicity or computational tractability, while giving rise to likely intrinsic errors that are difficult to detect.

Use of general vectorial feature representation assumes that measured features are significant for each class. Aggregation of information from different measurements contributes equally to the decisions. This can result in opacity in recognition mechanism (having no insight into decision processes) and misleading as well as inefficient classification because of the bluring effect on the discrimination power of each feature, especially when the patterns are highly context-sensitive and exhibit hierarchical structure. Furthermore this opacityness in recognition mechanism causes difficulty in incremental modification and debugging of the systems.

Inadequate use of the domain specific knowledge. Typically there is a lack of use of evidences from the domain constraints which are built up gradually during the recognition processes. In the case of chromosome classification for instance, expert heuristic knowledge can be modeled and formulated as contextual and/or relational constraints [14] (e.g. each class except class 24 (Y-chromosome) should normally contain at most two chromosomes; homologues should be similar; particular chromosome classes show particular characteristics in themselves or when compared with certain other chromosome classes, etc.). However such knowledge can not be incorporated in the above single-path decision theoretical scheme. There is a lack of flexible control mechanism in which hypothetical reasoning can be supported.

The points raised above indicate some of the major causes of the limitations of the conventional BPR approach. To overcome these limitations in order to improve the performance of existing BPR systems, a pilot case study has been carried out recently, in which we attempt to use a new scheme to achieve knowledge-based improvement of chromosome classification. This is described in the next section.

A knowledge-based BPR improvement scheme

Based on our study of KBS techniques and taking into account the problems encountered in BPR systems, a new scheme is developed for (but not restricted to) knowledge-based improvement of chromosome classification. By contrast to the single-path decision theoretical approach used in conventional BPR systems, a multistage recursive hypothesize-and-verify paradigm is adopted in such a scheme. The experience invested in classical statistical pattern recognition is retained, in that the existing pattern classifier is used to generate hypothetical classifications, and the domain expert knowledge is incorporated to form a 'constraint network', to help reduce the ambiguities which arise from the limitations of the existing techniques as mentioned in the previous section. Such a strategy is chosen on account that existing systems resulting from decades of statistical pattern recognition research and experience have already reached certain level of performance, and can be used to generate reasonably good hypotheses.

The 'constraint network' is a conceptual semantic net or relational graph we use to model the contextual and/or relational constraints formulated from the domain expert knowledge during his or her manual classification procedure. The constraint satisfaction is carried out by evidential reasoning in the hypothesis space of each pattern, in terms of rule-based updating of beliefs and verification of hypothetical classifications.

For the chromosome classification problem, a knowledge engineering effort has been made, in which we have attempted to acquire the domain knowledge from cytotechnicians, by interviewing and even learning to classify metaphase chromosomes.
Figure 3: An illustration of a 'constraint network'

ourselves [15]. The acquired knowledge we believe to be important in human expert classification procedure can be conceptualized and formulated in terms of domain constraints. These constraints can be categorized mainly into:

- Contextual constraints: e.g., in normal metaphase each class except class 24 (Y-chromosome) should contain at most two chromosomes; within each class two homologous chromosomes should be similar; the likelihood or confidence value to each class must be high; evidence from other cells in a multi-cell analysis, constraints from prior knowledge, etc.

- Relational constraints: e.g., relative size, relative number and contrast of the chromosome bands, particular chromosome classes show particular characteristics in themselves or when compared with certain other chromosome classes (chromosome No.13 has darker stain density below than above its midpoint; chromosome X has a larger centromeric index than chromosome No.7, etc.).

An illustration of such a 'constraint network' is depicted in Figure 3, in which the nodes represent different classes (or groups) to which chromosomes are assigned by hypothetical classifications, and the directional lines represent the (unary or n-ary) constraints forced on each of these class (or group) hypotheses. Since the verification of hypotheses in terms of constraint satisfaction is a highly pattern-driven inference process on its own, it is implemented by using production rules to match the symbolic descriptions of each chromosome and classification contexts.

Our scheme for knowledge-based improvement of chromosome classification consists of the following steps:

1. A classical pattern classifier of an existing chromosome analysis system is used to generate hypotheses and classification for each chromosome in the image under analysis. For each of these chromosomes a hypothesis space is constructed, and the computed a posteriori probabilities are used as the initial belief functions attributed to each of the singleton hypotheses. As chromosomes exhibit hierarchical pattern structures (Denver grouping), their hypothesis spaces are conveniently constructed into strictly hierarchical ones (fig.4). The group hypotheses denoted by $G_k, k = 1, ..., l$ are the subsets of singleton hypotheses denoted by $C_j, j = 1, ..., m$, and their initial belief values are simply the summations of the initial belief values of their children nodes, i.e.,

$$Bel(G_k) = \sum_{C_j \subseteq G_k} Bel(C_j)$$

2. Apply the Bayesian rule (maximum likelihood criterion) to obtain the hypothetical classification of each of the chromosomes. Now starts the verification of the hypothetical classifications. At this stage the constraint network in terms of a rule base is called in, to match against the symbolic descriptions of each of the chromosomes and classification contexts. The original beliefs residing in different nodes of the hypothesis space of each chromosome are updated using J. Pearl's evidential aggregation scheme [16]. This is carried out in the following way: For each hypothetical classification $H$ upon which the evidence bears directly, estimates the degree $\lambda_H$ to which the evidence confirms or disconfirms $H$. $\lambda_H$ is the likelihood ratio

$$\lambda_H = \frac{P(e|H)}{P(e|\neg H)}$$

Confirmation is expressed by $\lambda_H > 1$; disconfirmation by $\lambda_H < 1$. Such an estimation can be obtained by the subjective experience of the domain expert.

Each singleton hypothesis $C_j \in H$ obtains the weight $W_j = \lambda_H$ while every hypothesis outside $H$ receives a unity weight, $W_j = 1$. The belief in each singleton hypothesis $C_j$ is updated from the original value of $Bel(C_j)$ to:

$$Bel'(C_j) = P(C_j|e) = \alpha_H W_j Bel(C_j),$$

where $\alpha_H$ is a normalization factor:

$$\alpha_H = \left[ \sum_j W_j Bel(C_j) \right]^{-1}$$

The belief value of each intermediate-level hypothesis (i.e., group hypothesis) is computed by the summation of the belief values of its singleton elements (children class hypotheses).
has been wrongly classified into class 1 by a classical classifier in class 1 is displayed. The corresponding constraint rule discom- ments the initial hypothetical classification of this chromosome in fig.5(b), where the final classification result of that metaphase is shown mainly on account of availability, and integrating it with the conventional image processing and pattern recognition routines. During the experiments the estimations of odds were written so far, through our knowledge engineering attempt [15]. Experiments were carried out on a set of metaphase images with normal chromosome complement. The parameters initial and final misclassifications in Table.1 stand for the number of chromosomes that were misclassified in the initial classification by the classical classifier and in the final classification by applying the knowledge-based improvement scheme respectively. The misclassifications were determined by the human expert classifier examining at the pictorial outputs of the system. During the experiments the estimations of odds were based mainly on the domain knowledge acquired and empirical adjustment, a process in which tunings were frequently used. It was noted that the system performance could be affected by the chosen number of recursions (the chosen number was between 4 and 8 in the experiments), and by the choice of reject threshold which affected the rejection rate (we used a threshold of 0.3).

Conclusions

In order to overcome the limitations of the classical statistical pattern recognition strategy used in most BPR systems, and based on the fact that decision making in biomedical expert reasoning usually involves the use of heuristics from domain knowledge sources, a scheme was developed and proposed in this paper, for knowledge-based improvement of biomedical pattern recognition, particularly with respect to the classification strategy. Application of such a scheme has been attempted in a case study, to the field of automatic chromosome analysis which has been a hard BPR problem since long.

By contrast to the classification strategy used in BPR systems, in our scheme a multistage recursive hypothesize-and-verify paradigm is adopted. The experience invested in classical statistical pattern recognition is retained and used to generate hypothetical classifications, and the domain expert knowledge can be incorporated to form a 'constraint network', which is implemented in terms of rule-based updating of beliefs and verification of the hypothetical classifications.

The proposed scheme has been tested in our preliminary experiments with the metaphase chromosome image data. The results shown an encouraging improvement of classification performance by using such a scheme over the conventional BPR approach. This case study indicates that incorporating domain specific knowledge in BPR systems has considerable promise for success, though much work is still ahead towards the solution of the chromosome classification problem and a robust BPR paradigm.

Acknowledgements

The authors would like to thank Luc van Eycken for his assistance in \LaTeX{} coding of the manuscript adhering to the IEEE
CAIA specification. This work was supported by the Belgium National incentive program for fundamental research in Artificial Intelligence, initiated by the Belgium State - Prime Minister's Office - Science Policy programming, and by the Commission of the European Communities within the frame of its Medical and Health Research Program by project No. II.1.1/13. The scientific responsibility is assumed by its authors.

References


