Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 37(3), 2010, Pages 142–146
ISSN: 1223-6934

A k-nearest neighbor approach for chromosome shape classification

Mircea Sebastian Serbanescu

Abstract. The aim of this paper is to use a k-nearest neighbor algorithm for chromosome shape classification. To achieve this task a free web database of 117 normal karyotypes (for both men and women) was used. The classifier, in spite of its simplicity proved very reliable, providing a very good accuracy.

2010 Mathematics Subject Classification. Primary 68T10; Secondary 68T10.
Key words and phrases. chromosome classification, k-nearest neighbor, shape representation.

1. Introduction

Since Tjio and Levan discovered that the number of human chromosomes was 46 in 1956 [1] and the Denver group classification standard was established in 1960 [2], karyotyping of human chromosomes (Fig. 1) has became an important clinical procedure for screening and diagnosing genetic disorders and cancers [3]. Karyotyping is a standard technique utilized to classify metaphase chromosomes (46) into 24 types.

Manual karyotyping is a labor-intensive and time-consuming task (manual classification of 46 chromosomes in 24 classes done at least 20 times for one result), so developing automatic computer-assisted karyotyping systems has attracted significant research interest [4]. Fully automatic systems usually follow a number of consecutive steps [5]:

Received July 08, 2010. Revision received September 07, 2010.
The presented work has been conducted in the context of the GRANT-xy funded by The National University Research Council from Romania, under the contract 01/2008.

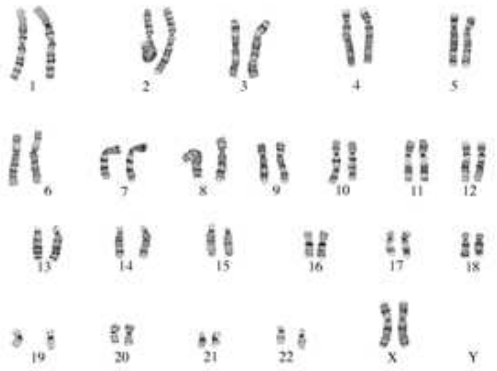

Figure 1. Normal kariotype, male

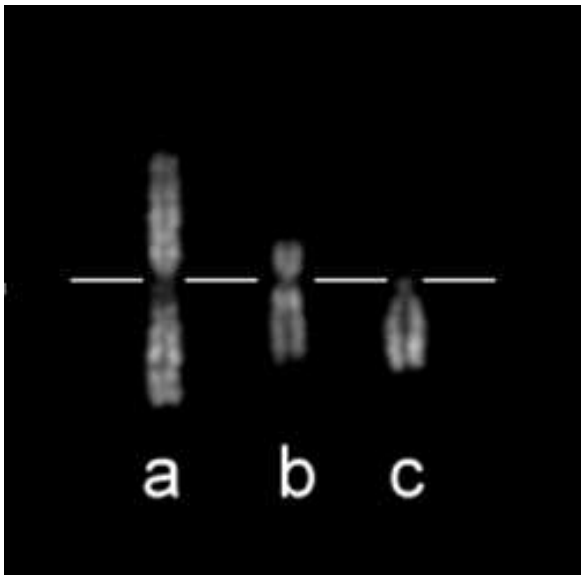

FIGURE 2. Chromosome shape types, a - metacentric, b - submetacentric, c - scrocentric

(1) *cleaning* of the image from stains and interphase nuclei;
(2) *segmentation* of the cleaned metaphase cell in its different chromosomes;
(3) extraction of *features* from all chromosomes;
(4) *classification* of the feature sets into the biological classes.

The first two steps employ methods from image processing, the last two from pattern recognition and statistics. Many classification methods have been tried in the aim of automated kariotyping, most of them using complex methods like: genetic algorithms [6], natural networks [4][7][8], but none of them offers a suitable solution in the matter of time and used resources.

The human achievement of classification uses three main features: size, shape and banding features.

2. Material and Method

Chromosome shapes are considered to be classified in three classes by taking in consideration the centromer position (the thinnest part of a chromosome): metacentric, submetacentric, scrocentric (Fig. 2).

The purpose of our paper was to correctly estimate the class of unknown chromosome, using a simple but efficient machine learning technique, namely the k-nearest neighbor algorithm.

The k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). The training objects are vectors in a multidimensional feature space, each with a class label. The training

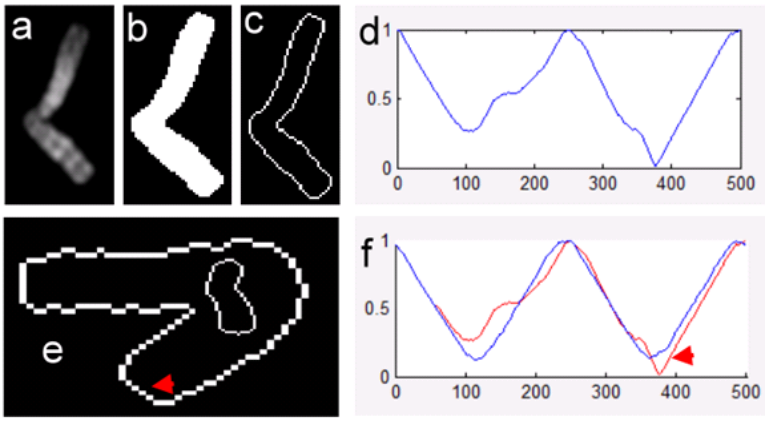

FIGURE 3. From chromosome picture to classifiable data. 3a - original chromosome, 3b - binarized image, 3c - bordered image, 3d - computed matrix, 3e matching chromosome shape (abstract view), 3f - maching different shapes.

phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point

For assessing the capability of this algorithm to correctly classify chromosomes, a free web database of 117 normal karyotypes (both men and women), consisting of 46 individual chromosomes (manually segmented by specialists) [9] was used in the aim of shape classification.

In order to obtain the border the images (Fig. 3a) had to be binarized (Fig. 3b). After the binarization possible inside objects were filled out. Next the one by one pixel connected border was extracted [12], resulting a matrix of coordinates (Fig. 3c). By estimating the distance between each pixel of the matrix and the center of gravity of the border we obtained a one dimension measurement charactering the shape of the chromosome. Last the measurement was resample at 500 points (in order to eliminate size matters) and was normalized to the maxim value(Fig. 3d).

The data was introduced in a k - nearest neighbor (k-NN) algorithm, implemented in Matlab (Mathworks) version 2007b. In order to compute the distance between different shapes the 500 points representation was interpreted as a circular list. Euclidian distance was used and the computed distances was obtained by shifting one border against the other until the best fit (minimum distance) (Fig. 3e and Fig.3f). The algorithm was run with different k values (number of neighbors taken in consideration for class estimation) varying from small values (k = 3 and 10) to larger values(k = 100).

3. Results

The results of 24 chromosomes from different classes (1-22,x,y) are presented in Table 1. The overall classification rate is good (around 88%) for a simple algorithm, but efficacious algorithm. Maximum classification rates were achieved for low values of the number of neighbors k. It is very interesting to observe that the best accuracy

TABLE 1. Classification results

	K=3	K=10	K=100
Misclassified (absolute value)	3	5	5
Correct classified (absolute value)	21	19	19
Classification accuracy (%)	87.5%	79.16%	79.61%

has been obtained for just k = 3 neighbors, proving its effectiveness. On the other hand, it is worth mentioning that up to 10 neighbors there is no improvement in the classification accuracy, the "elbow curve" showing a flat trend. Furthermore, this means that the in-class variation of shape is quite large.

On the other hand the algorithm is time consuming, average classification time is 12 seconds. Although the searching space is very big (5336 chromosomes) and for the computing of one distances there must be made 500 comparisons of 500 point matrix we must say this is not suitable for practical use.

The comparison to the results from other studies in literature is impossible. Some papers present multi-level classification methods, but none refers strictly to the shape classification. This is the originality of this attempt.

4. Conclusion

The overall result of chromosome shape classification using k-NN was good (approx. 90%). Higher classification rates were obtained for low values of the numbers of neighbors taken in consideration.

K-NN is a time consuming method and is not suitable for practical use, at least not for a large database like the one used in our paper.

5. Future work

In some future work we will try to reduce the searching space by estimating the most representing chromosome shapes of the classes and so reducing the number of comparisons or uses some synthetic descriptors of the shape, as Fourier descriptors, thus reducing the data dimensionality.

References

[1] J.H. Tjio and A. Levan, The chromosome number in man, *Hereditas* **42** (1956), 1–6.
[2] Conference D. A proposed standard system of nomenclature of human mitotic chromosomes. Lancet 1960; 1:1063-5.
[3] J. Piper, E. Granum, D. Rutovitz and H. Ruttledge, Automation of chromosome analysis, *Signal Process* **2** (1980), 203–221.
[4] X. Wang, B. Zheng, S. Li, J.J. Mulvihill, M.C. Wood and H. Liu, Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme.
[5] G. Ritter and L. Gao, Automatic segmentation of metaphase cells based on global context and variant analysis.
[6] J. Piper, Genetic algorithm for applying constraints in chromosome classification, *Pattern Recognition Letters* **16** (1995), 857–864.
[7] P.A. Errington and J. Graham, Application of artificial neural networks to chromosome classification, *Cytometry* **14**, no. 6, 627–639.
[8] W.P. Sweeney Jr., M.T. Musavi and J.N. Guidi, Classification of chromosomes using a probabilistic neural network, *Cytometry* **16**, no. 1, 17–24.

[9] E. Grisan, E. Poletti and A. Ruggeri, Automatic segmentation and disentangling of chromosome in Q-band prometaphase images, IEEE Trans Inf Technol B, in press, (2009).
[10] T.M. Cover and P.E. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* **13** (1967), no. 1, 21–27.
[11] P. Hall, B.U. Park and R.J. Samworth, Choice of neighbor order in nearest-neighbor classification, *Annals of Statistics* **36** (2008), 2135–2152.
[12] M.S. Serbanescu, Aria si perimetrul, informatie redundanta in contextul analizei cariotipului, *Zilele U.M.F. din Craiova*, Editura Medicala Universitara Craiova (2010), 1–45.

(Mircea Serbanescu) University of Medicine and Pharmacy of Craiova, Romania
E-mail address: mircea_serbanescu@yahoo.com