Outlier Treatment: A New Statistical Method for Automatic Chromosome Classification

María Teresa Gallegos and Gunter Ritter

Fakultät für Mathematik und Informatik, Universität Passau, 94030 Passau, Germany

Abstract. We design a new method for handling outliers in chromosome data for automatic classification in 24 classes. If, as proposed in Ritter et al., 1995, elliptically symmetric distributions are used our method reduces the cross-validation error by 50 % compared with the classical method of normal distributions without outlier treatment.

Keywords. Automatic chromosome classification, elliptically symmetric distributions, outliers, linear assignment problem

1 Automatic Classification of Chromosomes

A normal, nucleated, human cell contains 44 autosomal chromosomes and two sex chromosomes. The 44 autosomal chromosomes can be assigned to 22 classes 1..22 each of which consists of a matching pair of chromosomes. A representation of the chromosomal complement showing this class structure is called a *karyotype*. Attempts to automatically produce karyotypes from cellular images can be traced back to the 60s. This automation procedure needs methods of image analysis, statistical discriminance analysis, and combinatorial optimization. The image processing involves segmentation of the cellular image into its different chromosomes and feature extraction. We describe here how to effectively use these features for classifying the chromosomes into their 24 classes with the help of an outlier detection method.

2 Regular Observations and Outliers

Our method consists in deviding the complete population of each class into a *regular* and an *outlier* population. For doing this we use a variant of trimming, a known way of estimating parameters of populations contaminated by outliers (cf. Barnett and Lewis, 1994). Both populations are assumed to be elliptically symmetrically distributed (cf. Fang et al., 1990). An elliptically symmetric distribution is specified by its expectation e, covariance matrix V,

and radial function φ, its density being the function defined by

$$f(x) = det(V^{-1/2})\varphi(\|V^{-1/2}(x-e)\|);$$

here, $\|\cdot\|$ denotes Euclidean norm.

The parameters $e_{REG(j)}$, $V_{REG(j)}$, φ_{REG}, and $e_{OUT(j)}$, $V_{OUT(j)}$, φ_{OUT} of the regular and outlier populations are determined by the variant of the trimming method in the following way: Let us denote the whole, contaminated, population of class j by POP(j) and its expectation and variance by $e_{POP(j)}$ and $V_{POP(j)}$, respectively. An observation x is classified as regular in class j if its Mahalanobis distance from $e_{POP(j)}$ does not exceed a value cut_{REG} to be specified in advance (and independent of j). Denote the resulting regular population by REG(j) and its expectation and variance by $e_{REG(j)}$ and $V_{REG(j)}$, respectively. Now, an observation x is classified as an outlier from class j if its Mahalanobis distance from $e_{REG(j)}$ exceeds another value cut_{OUT}, the cutoff of the radial function φ_{REG} (i.e., $\varphi_{REG}(r) = 0$ for $r > cut_{OUT}$). The parameters of the outliers are $e_{OUT(j)}$, $V_{OUT(j)}$, and φ_{OUT}.

The minimization problem contained in the MAP-classifier over all possible assignments to quality (regular or outlier) and chromosomal classes (taking into account the correct number of chromosomes in each class) can be shown to be a linear assignment problem which can be efficiently solved. Indeed, writing c_{ij} for the minimum of the negative log-likelihood of chromosome i with respect to the two qualities of class j ($j \in 1..23$), the problem consists in minimizing the sum $\sum c_{ij}x_{ij}$ over all permutation matrices $(x_{ij}) \in (0..1)^{46\times 46}$ if it is known in advance that the cell is female. The transition to the general case (unknown sex, missing and extra chromosomes) is carried out by algebraic modifications.

Applied to the large Copenhagen data set Cpr, the method of outlier handling presented here reduces the cross-validation error rate by 25 % compared to the results of Ritter et al.,1995. Details will appear in a forthcoming paper.

References

Barnett V. and T. Lewis (1994). Outliers in statistical data. Wiley

Fang, K.-T, S. Kotz and K.-W. Ng (1990). Symmetric Multivariate and Related Distributions. Chapman and Hall, London, New York

Ritter, G., M.T. Gallegos, and K. Gaggermeier (1995). Automatic context-sensitive karyotyping of human chromosomes based on elliptically symmetric distributions. Pattern Recognition 28,6