

# Automated Knowledge Discovery using Parallel Coordinates

Alfred Inselberg\*  
Computer Science Department  
Tel Aviv University, Israel  
aiisreal@math.tau.ac.il

## Abstract

Classification is a basic task in Data Mining. Given a set  $S \subset P$ , a rule which distinguishes elements of  $S$  from those in  $P - S$  is sought. A **classifier** is an algorithm which outputs such rules. A measure of a classifier's efficacy is the error within which the rules it produces do the classification. Here two geometrically motivated **classifiers** based on Parallel Coordinates are presented. They are applied to real datasets, with both training and testing stages. The results have the least errors compared to those from 23 other classifiers. A dataset  $S$  involving  $N$  parameters is transformed to an  $N$ -dimensional point set. The algorithm builds a hypersurface which maximizes the number of points of  $S$  it contains and minimizes the number of points of  $P - S$  it excludes. The algorithm internally uses the representation in parallel coordinates of hypersurfaces it generates. Upon termination the hypersurface description is the classification rule. Two variants are considered, one called **Nested Cavities - NC** is composed of a cascade of decreasing cavities, while the other called **Enclosed Cavities - EC** consists of one set enclosing a number of non-intersecting cavities. The key properties are that either classifier :

- has very low computational complexity in the number of variables and the size of the dataset – contrasted with the very high or unknown (often unstated) complexity of other classifiers,
- the low complexity enables the rule derivation to be done in near real-time hence making the classification **adaptive** to changing conditions,
- provides comprehensible and explicit rules – contrasted to neural networks which act like “black boxes”,
- does dimensionality selection – where the minimal set of *original* variables (not transformed new variables as in Principal Component Analysis) required to state the rule is found,
- orders these variables so as to optimize the clarity of separation between the designated set and its complement – this solves the pesky “ordering problem” in parallel coordinates.

The algorithm is **display independent**, hence it can be applied to very large in size and number of variables datasets. Though it is instructive to present the results visually, the input size is no longer display-limited as for *visual* data mining.

Time permitting a new Decision Support System will also be presented. A multivariate relation, pertaining to a particular system, (i.e. the economy of a country) among  $N$  variables is represented and displayed as a hypersurface using parallel coordinates. Using an interior-point algorithm feasible states of the system (such as feasible economic policies) are interactively obtained and displayed showing salient *condition and constraint* dependent properties of the system. In this way, regions of sensitivity, effects *downstream* (i.e. on the remaining variables) of decisions as well as trade-off analysis can be rapidly performed and displayed.

---

\*Senior Fellow San Diego SuperComputing Center, and Multidimensional Graphs Ltd, Raanana, Israel

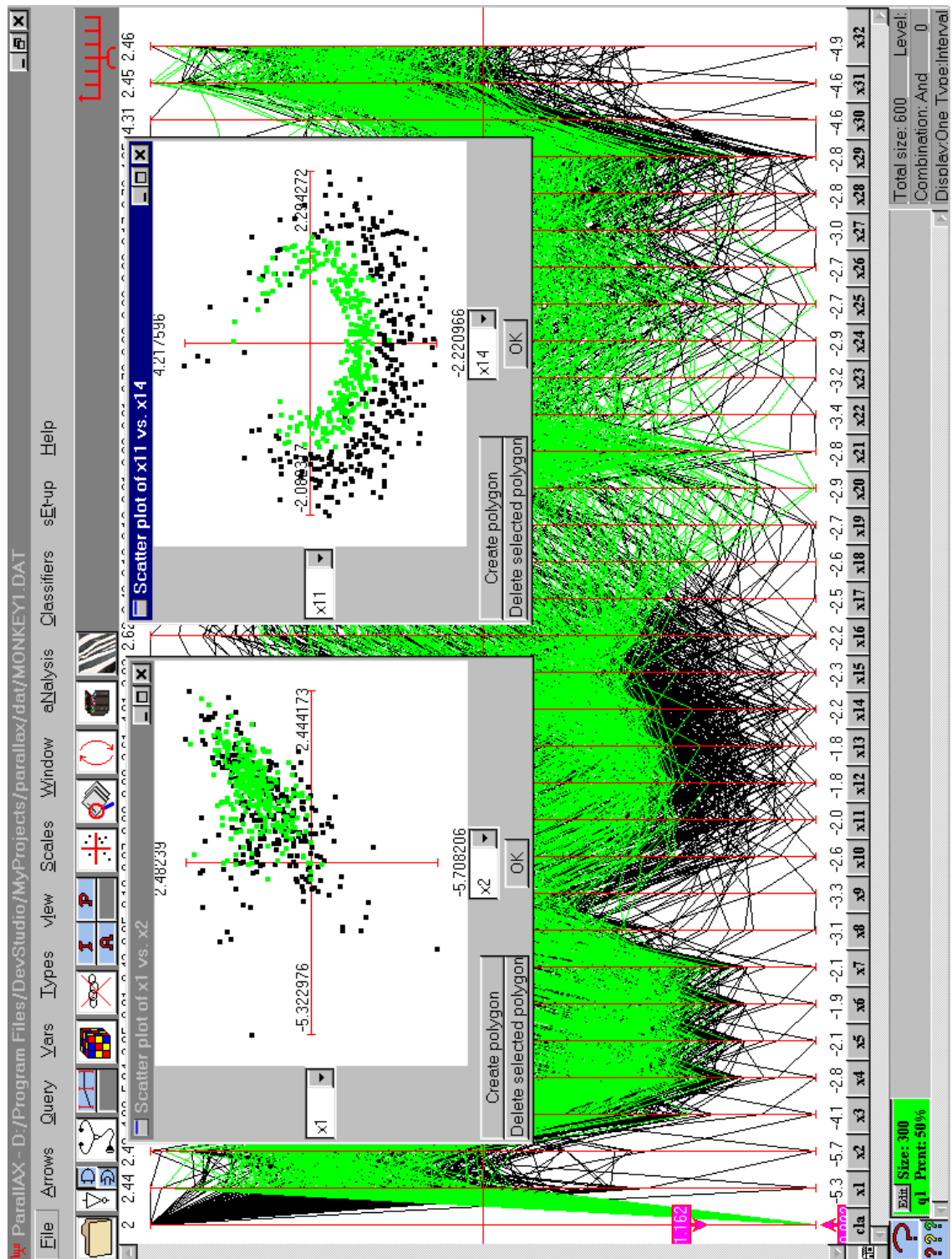


Figure 1: An example with measurements from 2 classes of neurons in the brain of a poor monkey involving 32 variables are shown in the background. A rule which distinguishes the outputs of the 2 neural classes is obtained using the NC classifier. The dimensionality selection found that only 9 of the variables are needed to make the distinction. The leftmost scatterplot is the plot of the first pair of variables in the original order with the data from the two classes intermingled. On the right is the first (i.e. “best”) pair of variables chosen by the dimensionality selection. Note the striking separation.