Chapter 7
Data Mining for Visualization of Health Processes (150 Patients)

General Purpose

Computer files of clinical data are often complex and multi-dimensional, and they are, frequently, hard to statistically test. Instead, visualization processes can be successfully used as an alternative approach to traditional statistical data analysis.

For example, Knime (Konstanz information miner) software has been developed by computer scientists from Silicon Valley in collaboration with technicians from Konstanz University at the Bodensee in Switzerland, and it pays particular attention to visual data analysis. It is used since 2006 as a freely available package through the Internet. So far, it is mainly used by chemists and pharmacists, but not by clinical investigators. This chapter is to assess, whether visual processing of clinical data may, sometimes, perform better than traditional statistical analysis.

Primary Scientific Question

Can visualization processes of clinical data provide insights that remained hidden with traditional statistical tests?

This chapter was previously published in "Machine learning in medicine-cookbook 3" as Chap. 1, 2014.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_7

Example

Four inflammatory markers (CRP (C-reactive protein), ESR (erythrocyte sedimentation rate), leucocyte count (leucos), and fibrinogen) were measured in 150 patients with pneumonia. Based on x-ray chest clinical severity was classified as A (mild infection), B (medium severity), C (severe infection). One scientific question was to assess whether the markers could adequately predict the severity of infection.

CRP	leucos	fibrinogen	ESR	x-ray severity
120,00	5,00	11,00	60,00	A
100,00	5,00	11,00	56,00	A
94,00	4,00	11,00	60,00	A
92,00	5,00	11,00	58,00	A
100,00	5,00	11,00	52,00	A
108,00	6,00	17,00	48,00	A
92,00	5,00	14,00	48,00	A
100,00	5,00	11,00	54,00	A
88,00	5,00	11,00	54,00	A
98,00	5,00	8,00	60,00	A
108,00	5,00	11,00	68,00	A
96,00	5,00	11,00	62,00	A
96,00	5,00	8,00	46,00	A
86,00	4,00	8,00	60,00	A
116,00	4,00	11,00	50,00	A
114,00	5,00	17,00	52,00	A

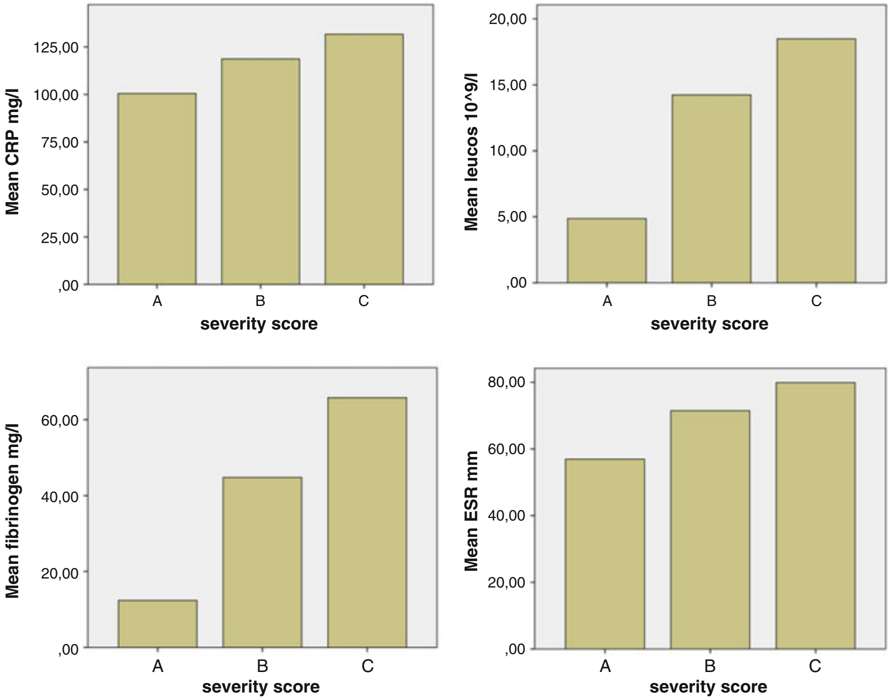
CRP = C-reactive protein (mg/l)
leucos = leucyte count (*10^9/l)
fibrinogen = fibrinogen level (mg/l)
ESR = erythrocyte sedimentation rate (mm)
x-ray severity = x-chest severity pneumonia score (A–C = mild to severe)

The data file is entitled “decisiontree”, and is available in extras.springer.com. Data analysis of these data in SPSS is rather limited. Start by opening the data file in SPSS statistical software.

Command:

click Graphs….Legacy Dialogs….Bar Charts….click Simple….click Define…. Category Axis: enter "severity score"….Variable: enter CRP….mark Other statistics….click OK.

After performing the same procedure for the other variables four graphs are produced as shown underneath. The mean levels of all of the inflammatory markers consistently tended to rise with increasing severities of infection. Univariate multinomial logistic regression with severity as outcome gives a significant effect of all of the markers. However, this effect is largely lost in the multiple multinomial logistic regression, probably due to interactions.



We are interested to explore these results for additional effects, for example, hidden data effects, like different predictive effects and frequency distributions for different subgroups. For that purpose Knime data miner will be applied. SPSS data files can not be downloaded directly in the Knime software, but excel files can, and SPSS data can be saved as an excel file (the csv file type available in your computer must be used).

Command in SPSS:

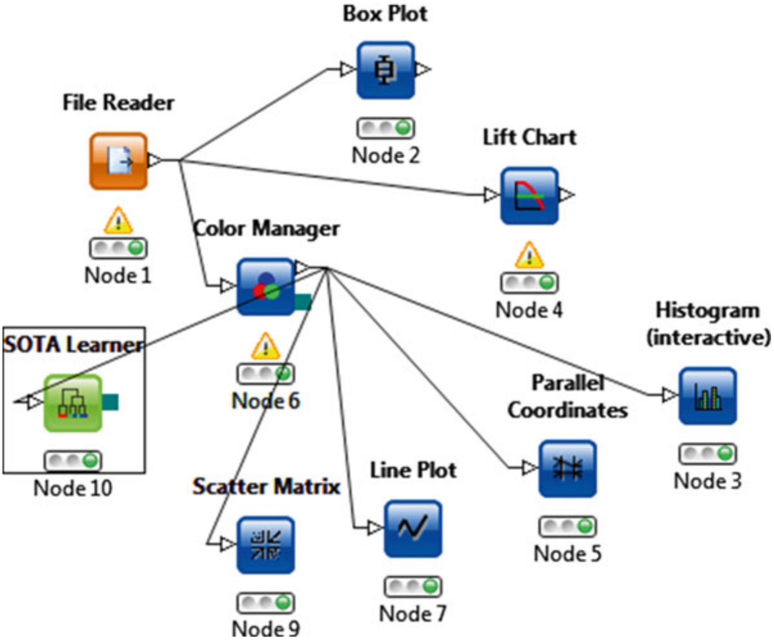
click File....click Save as....in "Save as" type: enter Comma Delimited (*.csv).... click Save.

Knime Data Miner

In Google enter the term “knime”. Click Download and follow instructions. After completing the pretty easy download procedure, open the knime workbench by clicking the knime welcome screen. The center of the screen displays the workflow editor like the canvas in SPSS modeler. It is empty, and can be used to build a stream of nodes, called workflow in knime. The node repository is in the left lower angle of the screen, and the nodes can be dragged to the workflow editor simply by left-clicking. The nodes are computer tools for data analysis like visualization and statistical processes. Node description is in the right upper angle of the screen. Before the nodes can be used, they have to be connected with the “file reader” node, and with one another by arrows drawn again simply by left clicking the small triangles attached to the nodes. Right clicking on the file reader enables to configure from your computer a requested data file....click Browse....and download from the appropriate folder a csv type Excel file. You are set for analysis now. For convenience an CSV file entitled “decisiontree” has been made available at extras.springer.com.

Knime Workflow

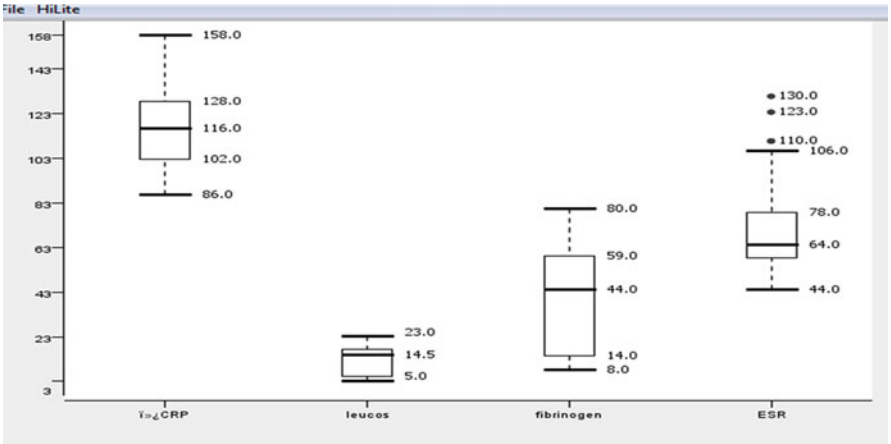
A knime workflow for the analysis of the above data example will be built, and the final result is shown in the underneath figure.



Box and Whiskers Plots

In the node repository find the node Box Plot. First click the IO option (import/export option nodes). Then click "Read", then the File Reader node is displayed, and can be dragged by left clicking to the workflow editor. Enter the requested data file as described above. A Node dialog is displayed underneath the node entitled Node 1. Its light is orange at this stage, and should turn green before it can be applied. If you right click the node's center, and then left click File Table a preview of the data is supplied.

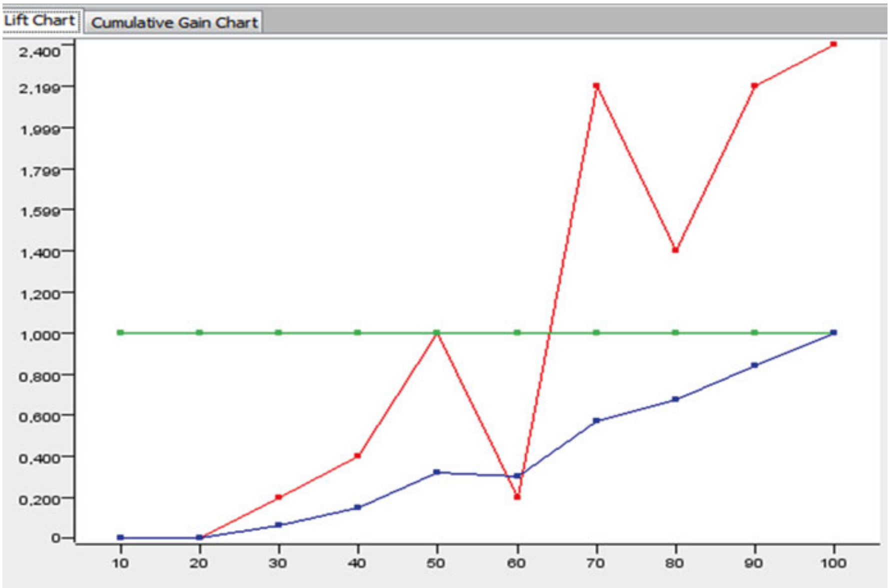
Now, in the search box of the node repository find and click Data Views....then "Box plot"....drag to workflow editor....connect with arrow to File reader....right click File reader....right click execute....right click Box Plot node....right click Configurate....right click Execute and open view....



The above box plots with 95 % confidence intervals of the four variable are displayed. The ESR plot shows that also outliers have been displayed The smallest confidence interval has the leucocyte count, and it may, thus, be the best predictor.

Lift Chart

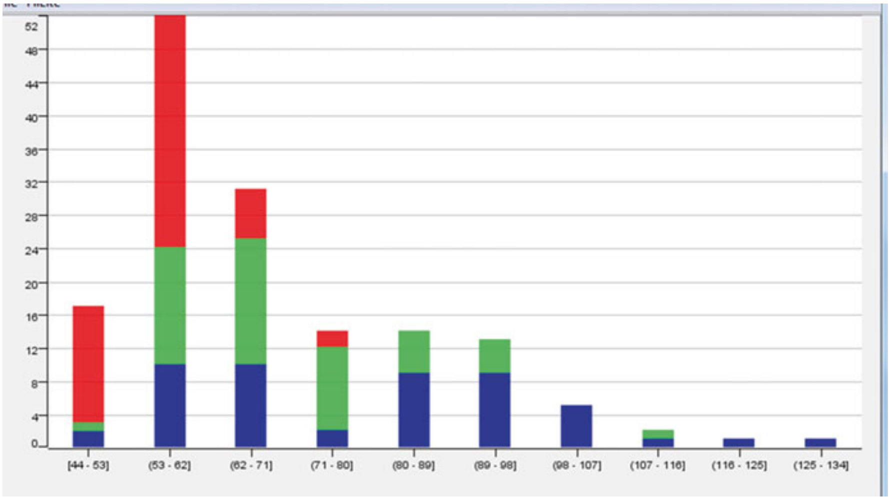
In the node repository....click Lift Chart and drag to workflow editor.... connect with arrow to File reader....right click execute Lift Chart node....right click Configurate....right click Execute and open view....



The lift chart shows the predictive performance of the data assuming that the four inflammatory markers are predictors and the severity score is the outcome. If the predictive performance is no better than random, the ratio successful prediction with/without the model = 1.000 (the green line) The x-axis give dociles (1 = 10 = 10 % of the entire sample etc.). It can be observed that at 7 or more dociles the predictive performance start to be pretty good (with ratios of 2.100–2.400). Logistic regression (here multinomial logistic regression) is being used by Knime for making predictions.

Histogram

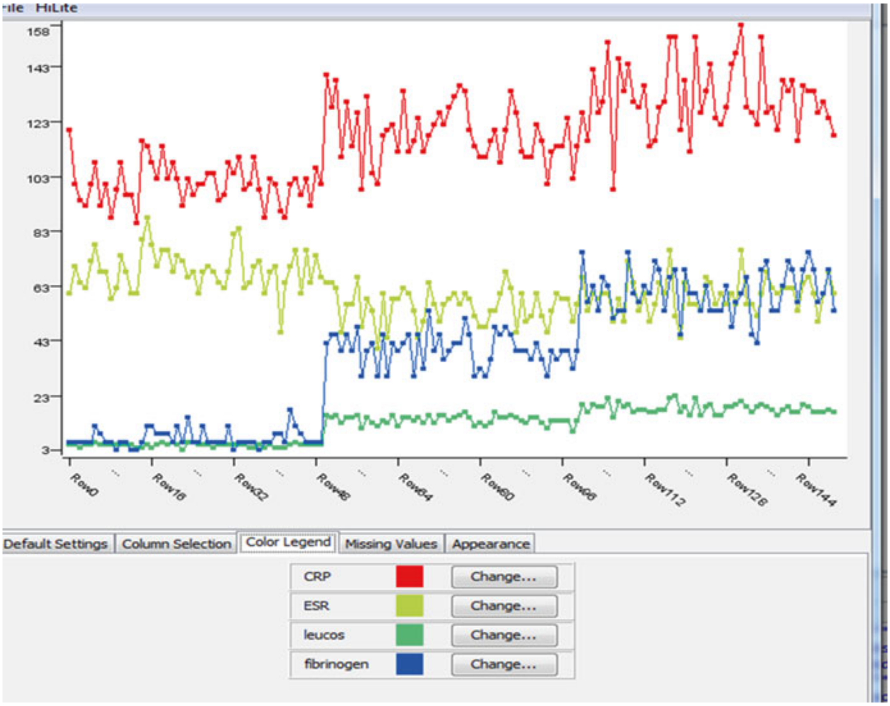
In the node repository click type color....click the color manager node and drag to workflow editor....in node repository click color....click the Esc button of your computer....click Data Views....select interactive histogram and transfer to workflow editor....connect color manager node with File Reader…connect color manager with “interactive histogram node”....right click Configurate....right click Execute and open view....



Interactive histograms with bins of ESR values are given. The colors provide the proportions of cases with mild severity (A, red), medium severity (B, green), and severe pneumonias (C, blue). It can be observed that many mild cases (red) are in the ESR 44–71 mm cut-off. Above ESR of 80 mm blue (severe pneumonia) is increasingly present. The software program has selected only the ESR values 44–134. Instead of histograms with ESR, those with other predictor variables can be made.

Line Plot

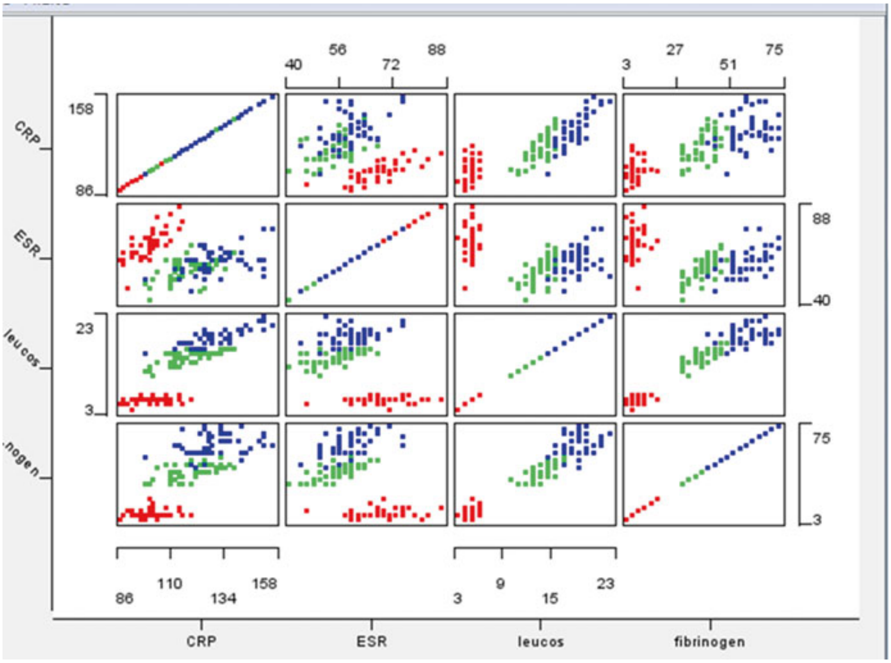
In the node repository click Data Views....select the node Line plots and transfer to workflow editor....connect color manager with "Line plots"....right click Configurate....right click Execute and open view....



The line plot gives the values of all cases along the x-axis. The upper curve are the CRP values, The middle one the ESR values. The lower part are the leucos and fibrinogen values. The rows 0–50 are the cases with mild pneumonia, the rows 51–100 the medium severity cases, and the rows 101–150 the severe cases. It can be observed that particularly the CRP-, fibrinogen-, and leucos levels increase with increased severity of infection. This is not observed with the ESR levels.

Matrix of Scatter Plots

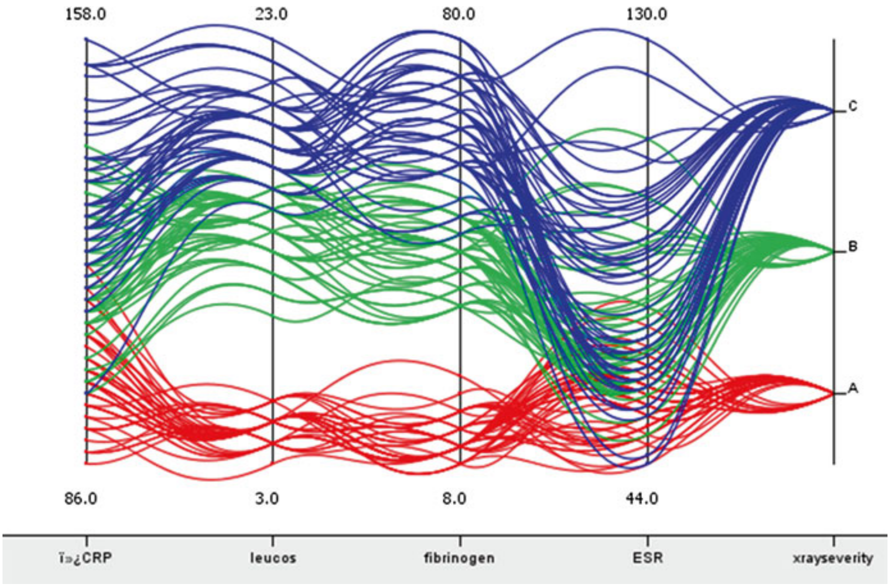
In the node repository click Data Views....select “Matrix of scatter plots” and transfer to workflow editor....connect color manager with “Matrix of scatter plots” .... right click Configurate....right click Execute and open view....



The above figure gives the results. The four predictors variables are plotted against one another. by the colors (blue for severest, red for mildest pneumonias) the fields show that the severest pneumonias are predominantly in the right upper quadrant, the mildest in the left lower quadrant.

Parallel Coordinates

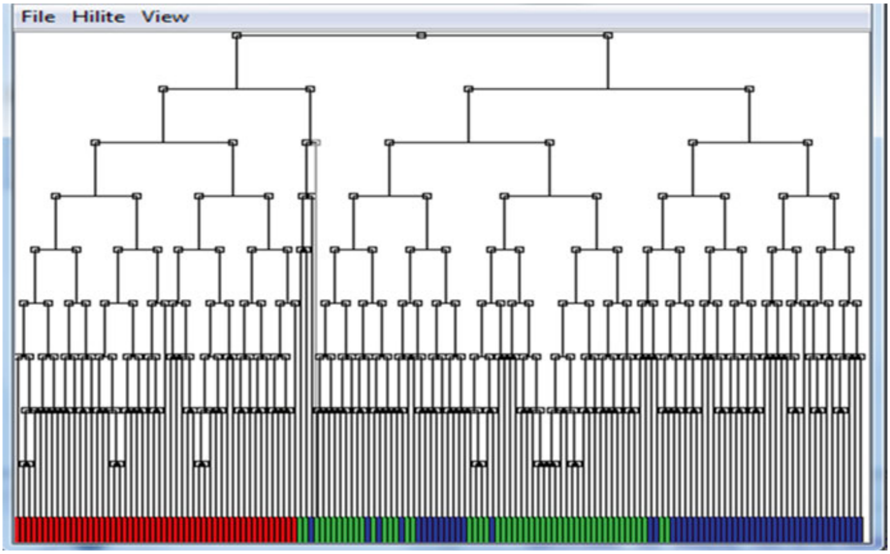
In the node repository click Data Views....select “Parallel coordinates” and transfer to workflow editor....connect color manager with “Parallel coordinates” ....right click Configurate....right click Execute and open view....click Appearance....click Draw (spline) Curves instead of lines....



The above figure is given. It shows that the leucocyte count and fibrinogen level are excellent predictors of infection severities. CRP and ESR are also adequate predictors of infections with mild and medium severities, however, poor predictors of levels of severe infections.

Hierarchical Cluster Analysis with SOTA (Self Organizing Tree Algorithm)

In the node repository click Mining....select the node SOTA (Self Organizing tree Algorithm) Learner and transfer to workflow editor....connect color manager with “SOTA learner”....right click Configurate....right click Execute and open view....



SOTA learning is a modified hierarchical cluster analysis, and it uses in this example the between-case distances of fibrinogen as variable. On the y-axis the standardized distances of the cluster combinations. Clicking the small squares interactively demonstrates the row numbers of the individual cases. It can be observed at the bottom of the figure that the severity classes very well cluster, with the mild cases (red) left, medium severity (green) in the middle, and severe cases (blue) right.

Conclusion

Clinical computer files are complex, and hard to statistically test. Instead, visualization processes can be successfully used as an alternative approach to traditional statistical data analysis. For example, Knime (Konstanz information miner) software developed by computer scientists at Konstanz University Technical Department at the Bodensee, although mainly used by chemists and pharmacists, is able to visualize multidimensional clinical data, and this approach may, sometimes, perform better than traditional statistical testing. In the current example it was able to demonstrate the clustering of inflammatory markers to identify different classes of pneumonia severity. Also to demonstrate that leucocyte count and fibrinogen were the best markers, and that ESR was a poor marker. In all of the markers the best predictive performance was obtained in the severest cases of disease. All of these observations were unobserved in the traditional statistical analysis in SPSS.

Note

More background, theoretical and mathematical information of splines and hierarchical cluster modeling are in Machine learning in medicine part one, Chap. 11, Non-linear modeling, pp 127–143, and Chap. 15, Hierarchical cluster analysis for unsupervised data, pp 183–195, Springer Heidelberg Germany, from the same authors.