Journal of Biomedical Informatics 66 (2017) 95–104

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Machine-learned cluster identification in high-dimensional data

Alfred Ultsch [a], Jörn Lötsch [b,c,*]

[a] DataBionics Research Group, University of Marburg, Hans-Meerwein-Straβe, 35032 Marburg, Germany
[b] Institute of Clinical Pharmacology, Goethe-University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany
[c] Fraunhofer Institute of Molecular Biology and Applied Ecology-Project Group Translational Medicine and Pharmacology (IME-TMP), Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

ARTICLE INFO

Article history:
Received 14 August 2016
Revised 27 November 2016
Accepted 26 December 2016
Available online 28 December 2016

Keywords:
Machine-learning
Clustering

ABSTRACT

Background: High-dimensional biomedical data are frequently clustered to identify subgroup structures pointing at distinct disease subtypes. It is crucial that the used cluster algorithm works correctly. However, by imposing a predefined shape on the clusters, classical algorithms occasionally suggest a cluster structure in homogenously distributed data or assign data points to incorrect clusters. We analyzed whether this can be avoided by using emergent self-organizing feature maps (ESOM).
Methods: Data sets with different degrees of complexity were submitted to ESOM analysis with large numbers of neurons, using an interactive R-based bioinformatics tool. On top of the trained ESOM the distance structure in the high dimensional feature space was visualized in the form of a so-called U-matrix. Clustering results were compared with those provided by classical common cluster algorithms including single linkage, Ward and k-means.
Results: Ward clustering imposed cluster structures on cluster-less "golf ball", "cuboid" and "S-shaped" data sets that contained no structure at all (random data). Ward clustering also imposed structures on permuted real world data sets. By contrast, the ESOM/U-matrix approach correctly found that these data contain no cluster structure. However, ESOM/U-matrix was correct in identifying clusters in biomedical data truly containing subgroups. It was always correct in cluster structure identification in further canonical artificial data. Using intentionally simple data sets, it is shown that popular clustering algorithms typically used for biomedical data sets may fail to cluster data correctly, suggesting that they are also likely to perform erroneously on high dimensional biomedical data.
Conclusions: The present analyses emphasized that generally established classical hierarchical clustering algorithms carry a considerable tendency to produce erroneous results. By contrast, unsupervised machine-learned analysis of cluster structures, applied using the ESOM/U-matrix method, is a viable, unbiased method to identify true clusters in the high-dimensional space of complex data.

1. Introduction

High-dimensional data is increasingly generated in biomedical research. An intuitive approach at utilizing these data is the search for structures such as the organization into distinct clusters. For example, gene expression profiling by grouping genes and samples simultaneously is a widespread practice used to identify distinct subtypes of diseases [1–3]. Usually, disease-specific expression-patterns are displayed on a clustered heatmap [4] as the most popular graphical representation of high dimensional genomic data [5]. Such plots show the cluster, respectively distance, structure at the margin of the heatmap as a dendrogram. A typical example result of this approach is shown in Fig. 1 that resembles results of genetic profiling analyses where several subgroups were suggested [2,3,6,7].

However, the data underlying the heatmap in Fig. 1 is displayed in Fig. 2. It consists of an artificial data set with 4002 points, in a 3D view resembling a golf ball [8], that with its equidistant distance distribution lacks any cluster structure. The apparent structure seen in the heatmap of Fig. 1 (left panel) is a direct result from a weakness of most clustering algorithms. That is, these methods impose a structure onto the data instead of identifying structure in the data. The majority of clustering algorithms use an implicit or explicit shape model for the structure of a cluster, such as a sphere in *k*-means or a hyperellipsoid in Ward clustering. This means, given a predefined number of clusters *k*, a clustering

* Corresponding author at: Goethe-University, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany.

E-mail address: j.loetsch@em.uni-frankfurt.de (J. Lötsch).

http://dx.doi.org/10.1016/j.jbi.2016.12.011
1532-0464/© 2017 The Authors. Published by Elsevier Inc.

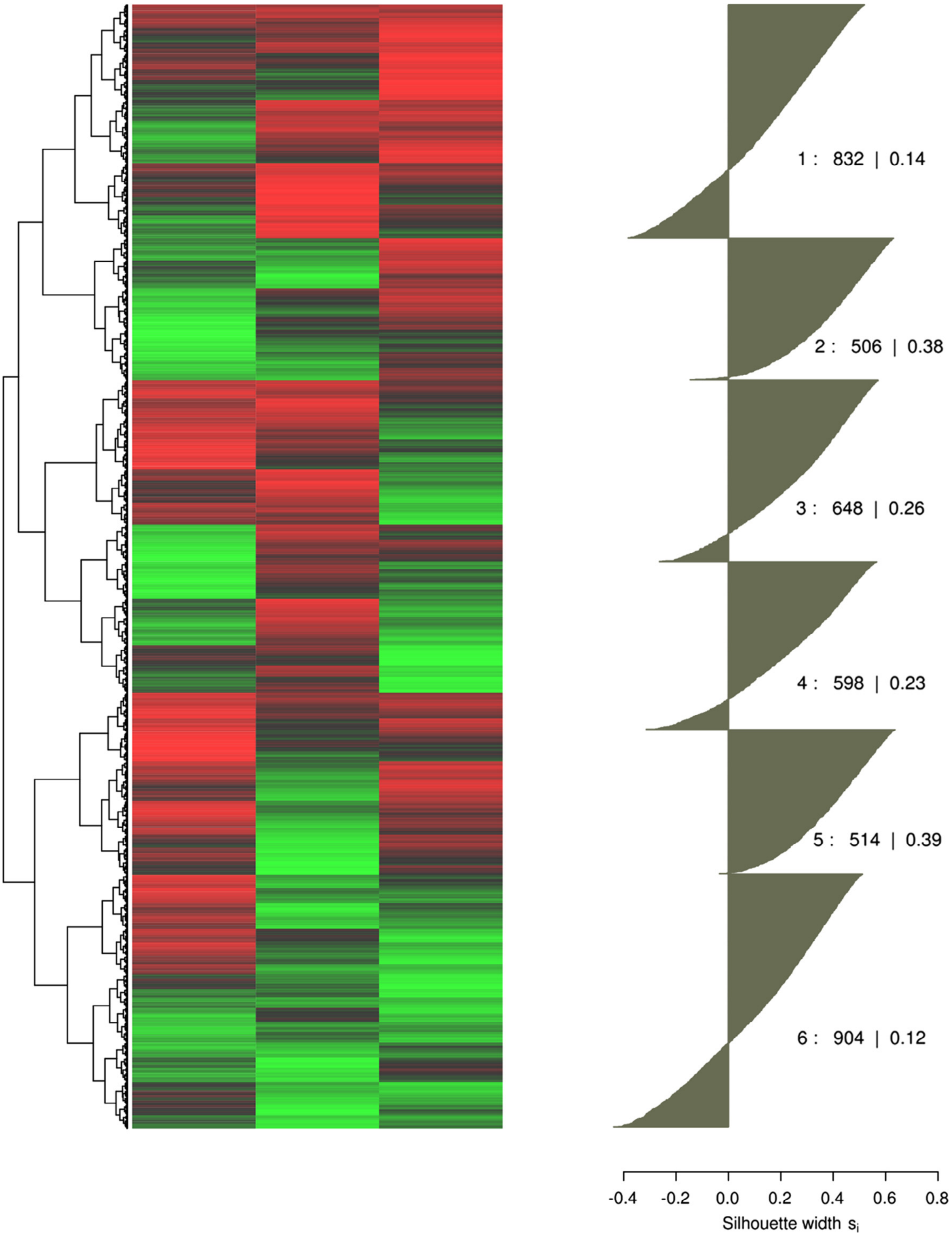

Fig. 1. Visualization of high dimensional data. Left panel: data presented in the form of a clustered heatmap as commonly used to identify groups of subjects (rows) sharing a similar gene expression profile. Data is presented color coded with smaller values in red and larger values in green. The dendrogram at the left margin of the matrixplot shows the hierarchical cluster structure. This suggests several distinct clusters up to possibly 4–11 subgroups, for which the right panel shows a silhouette plot for a six cluster solution. The silhouette coefficients for the six clusters indicate how near each sample is to its own relative to neighboring clusters. Values near +1 indicate that the sample is far away from the neighboring clusters while negative values indicate that those samples might have been assigned to the wrong cluster. The average silhouette coefficient is positive. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

algorithm calculates the coverage of the data with *k* of these geometric shapes, independently of whether or not this fits the structure of the data. This can result in erroneous cluster associations of samples or in the imposing of cluster structures non-existent in the data.

The example (Figs. 1 and 2) shows how cluster algorithms may suggest a more complex data structure than truly present. Clustering algorithms such as those mentioned above are implemented in standard data analysis software packed with laboratory equipment or in widely used statistical data analysis software packages.