Outlier Detection through Bipartite Visual Analytics

Suresh K. Bhavnani[1] PhD, Justin A. Drake[1] BS, Bryant Dang[1] BS, Shyam Vishweswaran[2], MD PhD
[1]Inst. for Translational Sciences, UTMB; [2]Dept. of Biomedical Informatics, University of Pittsburgh

Abstract

A critical goal of outlier detection is to determine whether an outlying value was caused by experimental/human error, or by natural biological diversity. However, because univariate or multivariate methods (e.g., box plots and principle component analysis) typically used for outlier detection use unipartite representations, they cannot distinguish whether outliers across a set of variables represent, for example, a single patient or different patients. Here we propose a bipartite visual analytical approach to outlier detection, and demonstrate its usefulness for identifying complex bipartite outliers in a dataset of rickettsioses patients, which enabled domain experts to determine whether the outliers were caused by errors, or by biological diversity.

Introduction

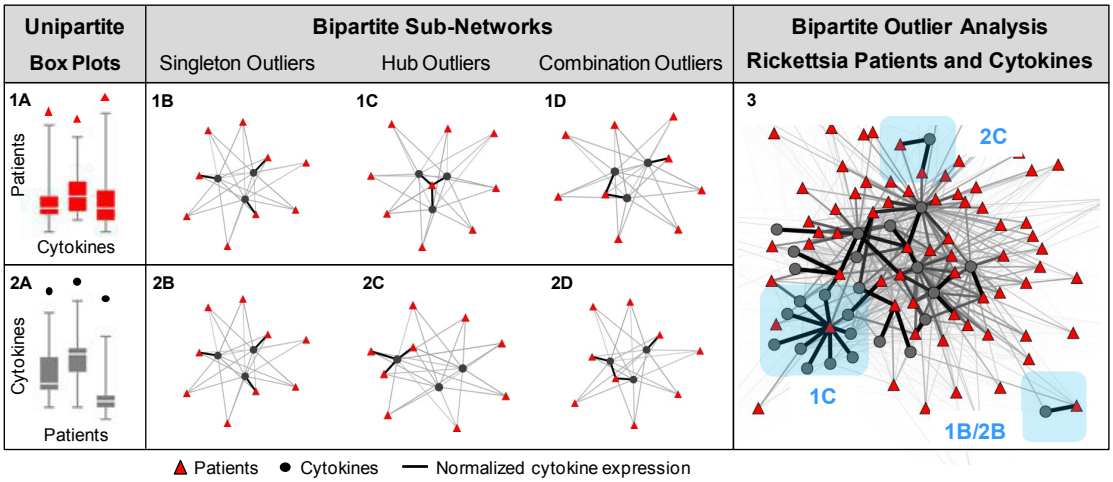
Comprehending the nature of outliers can help domain experts determine whether the outlying values were caused by experimental/human error or by natural biological diversity[1], a decision which can profoundly affect results and their interpretation. However, current methods typically used for outlier detection do not provide a bipartite view of the outliers, potentially concealing important domain information. For example, as shown in Figure 1A, box plots are typically used to show how far outliers (e.g., red triangles representing patients) are from the mean or median of each variable (e.g., cytokines). However, such unipartite approaches cannot easily reveal whether the three outliers shown represent the same or different patients. We therefore posed the research question: *Can a bipartite visual analytical representation help to distinguish between different types of bipartite outliers?*

Method

We developed a framework of outliers which compared box plots to bipartite networks. Nodes in the networks represented patients and cytokines, and edges between nodes represented normalized cytokine expression values. The framework was used to identify outliers in a real dataset[2], which were then interpreted by a domain expert.

Results and Conclusion

As shown in Figure 1, the unipartite box plots preserved the identity of either the cytokines (1A), or the patients (2A), but not both. In contrast, the bipartite networks preserved the identity of both entities in the data, revealing whether the outliers represented *singleton outliers* (1B, 2B), *hub outliers* (1C, 2C), or *combination outliers* (1D, 2D) of either patients or cytokines. This framework helped to identify three bipartite outliers (confirmed by Grubb's test at the .05 level) in a dataset of rickettsioses patients and cytokine expressions which were each determined to be biologically relevant to the study based on their bipartite relationships. Future research should test the usefulness of the framework on other datasets, and its application to other bipartite representations such as Circos ideograms.



Figures 1-3. Unipartite box plots (1A, 2A) are not designed to disambiguate between bipartite outliers (1B-1D, 2B-2D), which were used to identify outliers and their causes in a data set of Rickettsia patients and cytokines (3). Funded in part by IHII, & CDC R21OH009441-01A2.

References

1. Motulsky, H. *Intuitive Biostatistics*. Oxford University Press. 1st edition. 1995.
2. Bhavnani S.K., et al. How Cytokines Co-occur across Rickettsioses Patients: From Bipartite Visual Analytics to Mechanistic Inferences of a Cytokine Storm. *AMIA Summit on Translational Bioinformatics* (in press).