Nature Inspired Visualization of Unstructured Big Data

Aaditya Prakash

prakash@aaditya.info

Motivation

- Unstructured data is ubiquitous and is assumed to be around 80% of all data generated[1]
- Lack of recognizable structure and huge size makes it very challenging to work with Unstructured Large Datasets
- Classical Visualization – not suited for BigData; slow, memory hogging, limited dimensions

Self Organizing Maps

- Unsupervised Machine Learning Technique
- Provides dimension reduction
- Multivariate Analysis
- Fast and low on memory (2D planar images)
- Reconstructing Self Organizing Maps as Spider Graphs for better visual interpretation

[1] Unstructured Data and the 80 Percent Rule, Clarabridge Bridgepoints, 2008 Q3. http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551

Self Organizing Maps

- Artificial Neural Networks proposed by Teuvo Kohonen[1] which transforms the input dataset into two dimensional lattice
- Points in input layer are mapped onto 2D lattice, making each point potentially a Neuron

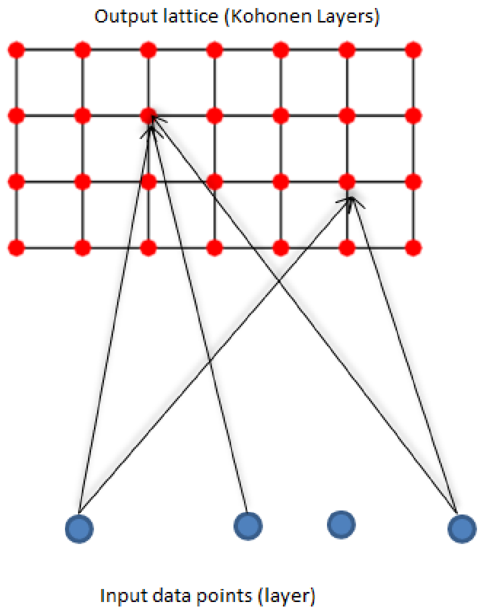


Figure: Kohonen Network

$$d_j(\mathbf{x}) = \sum_{i=1}^{D} (x_i - w_{ji})^2$$

Figure: Discriminant Function

where,

x = point on Input Layer

w = weight of the input point (x)

i = all the input points

j = all the neurons on the lattice

d = Euclidean distance

[1] Kohonen, T.; "The self-organizing map," Proceedings of the IEEE , vol.78, no.9 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=58325&isnumber=21

Current Visualization of SOM

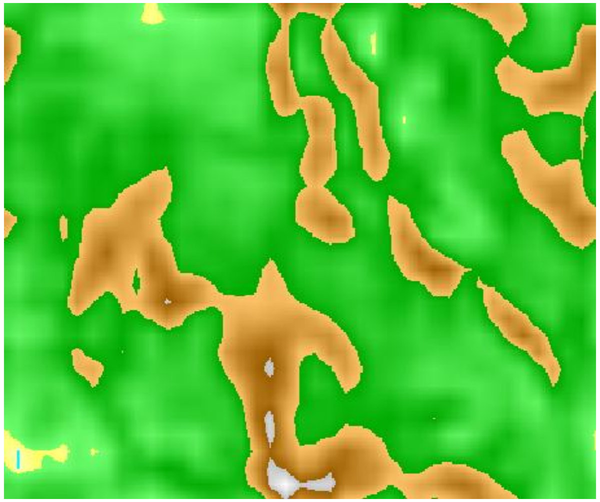


Fig: RapidMiner Tool (AGPL)

Shows the Kohonen Map obtained after training the Neurons

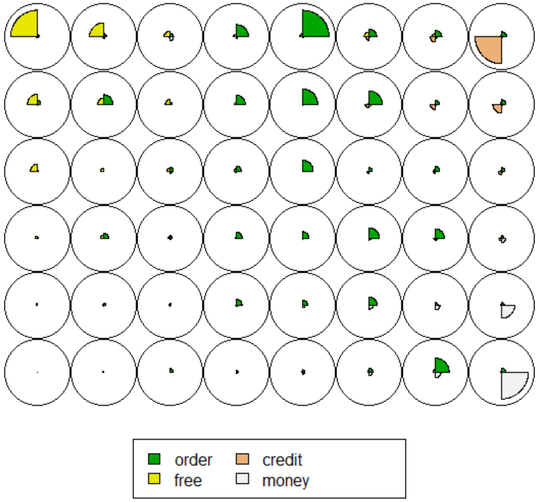


Fig: ‘R’, package ‘Kohonen’

Intervariate plot of 4 frequent words in Spam

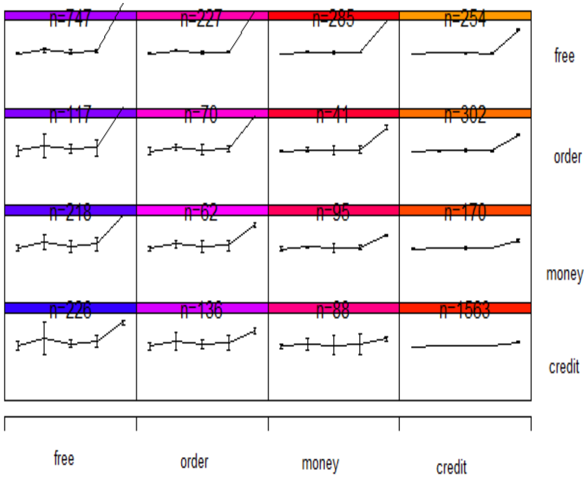


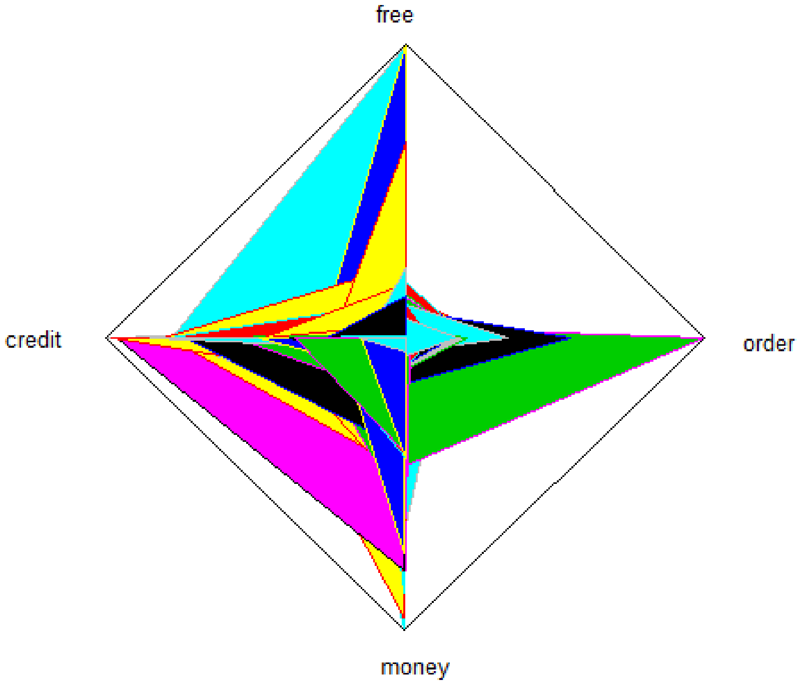
Fig: ‘R’, package ‘SOM’

Regression of the same four words

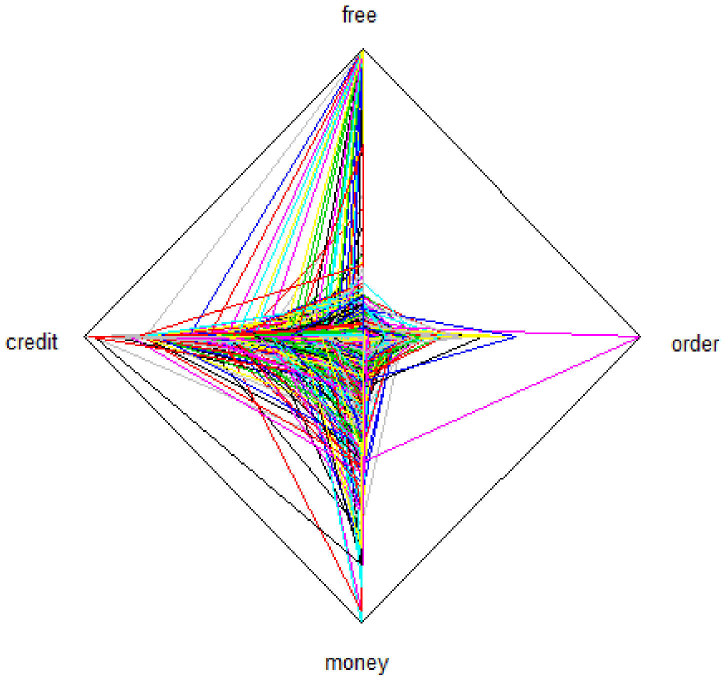
Algorithm

1. Filter the results
2. Make a polygon with as many sides as the variables.
3. Make the radius of the polygon to be the maximum of the value in the dataset.
4. Draw the grid for the polygon
5. Make segments inside the polygon if the strength of the two variables inside the segment is greater than the specified threshold.
6. Loop Step 5 for every variable against every other variable
7. Color the segments based on the frequency of variable.
8. Color the line segments based on the threshold of each variable pair plotted.

Spider Plots

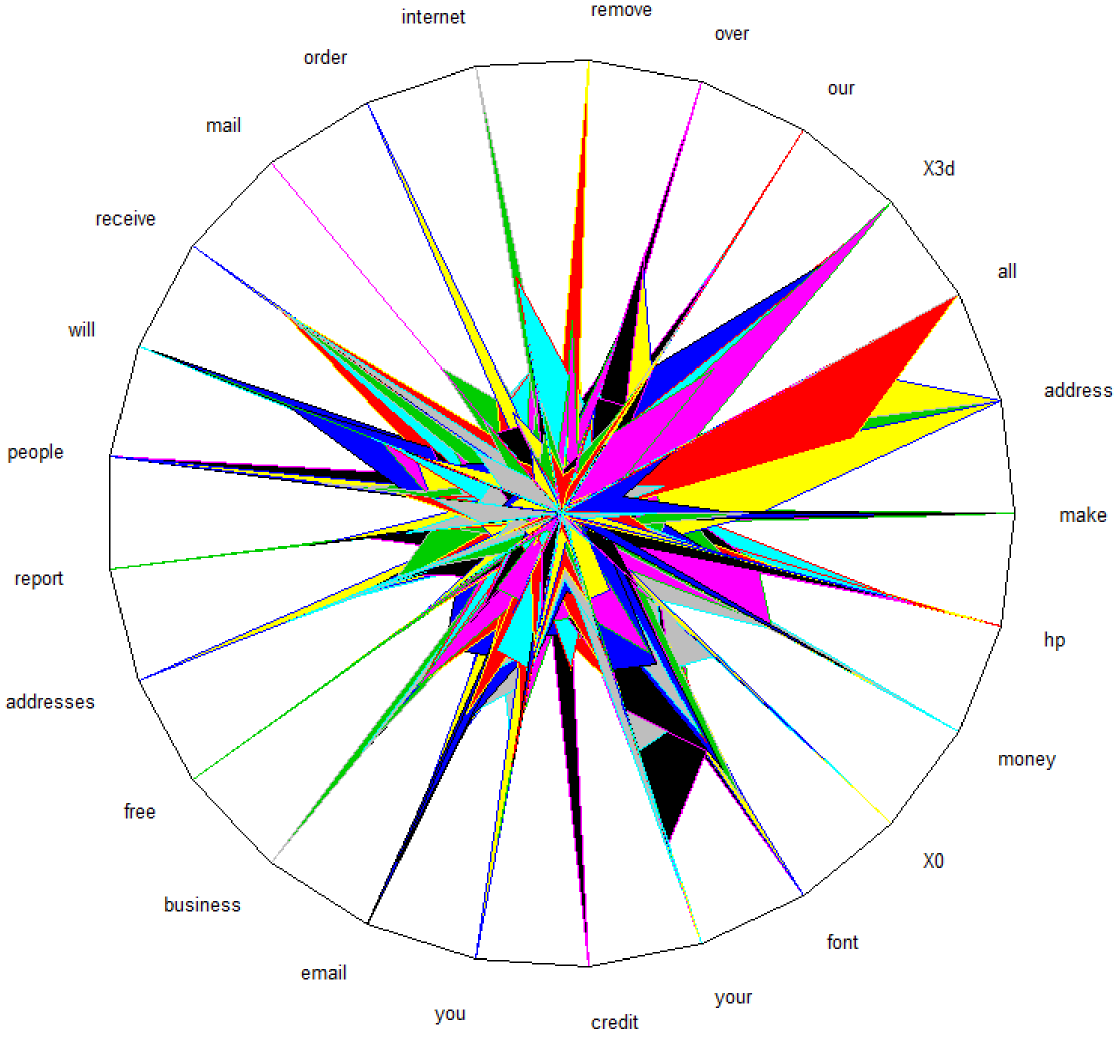


SOM visualization in R using the Algorithm given above. (showing segments i.e inter-variable dependency)



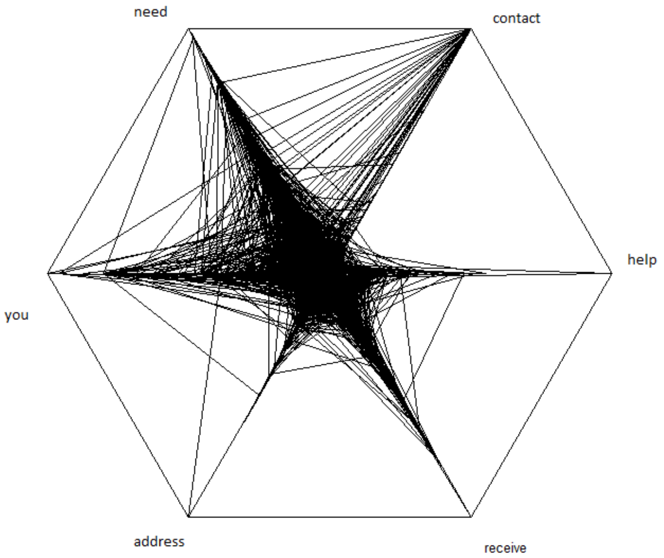
SOM visualization in R using Algorithm given above (showing threads, i.e inter-variable strength)

Big picture for Big Data



Conclusion

- Analyze inter-variate relations
- No need to convert Unstructured to Structured Data
- Advantages of Machine Learning and Visualization in single step
- Discover hidden relationships and potentially mining oppurtunities



Scope

- Enhance to work with images, sound and videos
- Dynamic representation to show live changes

References

- Kohonen, T.; "The self-organizing map," Proceedings of the IEEE , vol.78, no.9, pp.1464-1480, Sep 1990
- Teuvo Kohonen, Panu Somervuo, How to make large self-organizing maps for nonvectorial data, Neural Networks, Volume 15, Issues 8–9, October–November 2002
- Gail A. Carpenter, Stephen Grossberg, A massively parallel architecture for a self-organizing neural pattern recognition machine, Computer Vision, Graphics, and Image Processing, Volume 37, Issue 1, January 1987, Pages 54-115
- R. Wehrens and L.M.C. Buydens, Self- and Super-organising Maps in R: the kohonen package J. Stat. Softw., 21(5), 2007
- Jun Yan, Self-Organizing Map (with application in gene clustering) in R
- Gordon V. Cormack. 2008. Email Spam Filtering: A Systematic Review. *Found. Trends Inf. Retr.*1, 4 (April 2008)
- Anurat Chapanond, Mukkai S. Krishnamoorthy, and B\&\#252;lent Yener. 2005. Graph Theoretic and Spectral Analysis of Enron Email