Exploring high dimensional data with Butterfly: a novel classification algorithm based on discrete dynamical systems

Joseph Geraci, Moyez Dharsee, Paulo Nuin, Alexandria Haslehurst, Madhuri Koti, Harriet E. Feilloter and Ken Evans

Bioinformatics, vol.30 no.5 2014, pages 712-718

Fanny-Dhelia Pajuste

JC 21.03.2014

Motivation

- Visualizing high dimensional data using discrete dynamical system
- 2D representation of the relationships between subjects without geometric projections, transformed axes or principal components
- Human readable representation of data
- Detecting unrevealed clusters

Introduction (1)

- *n*-dimensional geometry (*n* variables)
- Geometry changes based on what variables we include/emphasize
- Selected variables determine the relationships via the distance
- How to study possible variable sets and corresponding relationships in an efficient, accurate and reproducable way?

Introduction (2)

- Reveals subclusters even in apparently homogeneous main clusters
- Bottom up exploration of data in addition to a procedure for testing ang training protocol
- Many tools overfit the data due to the presence of noise
- Data points are tranformed into a 2D space
- Clusters are separated by straight lines
- Non-linear correlations are identified between features
- The relationships in space remain simple

Introduction (3)

- They introduce Butterfly and it efficacy in handling molecular profiling datasets
- It is presented as a data exploration and machine learning classification tool
- Applied to high dimensional data after some feature reduction step using 50 or less dimensions
- They only introduce the algorithm and its ability to classify data
- They don’t compare it with other tools

Background

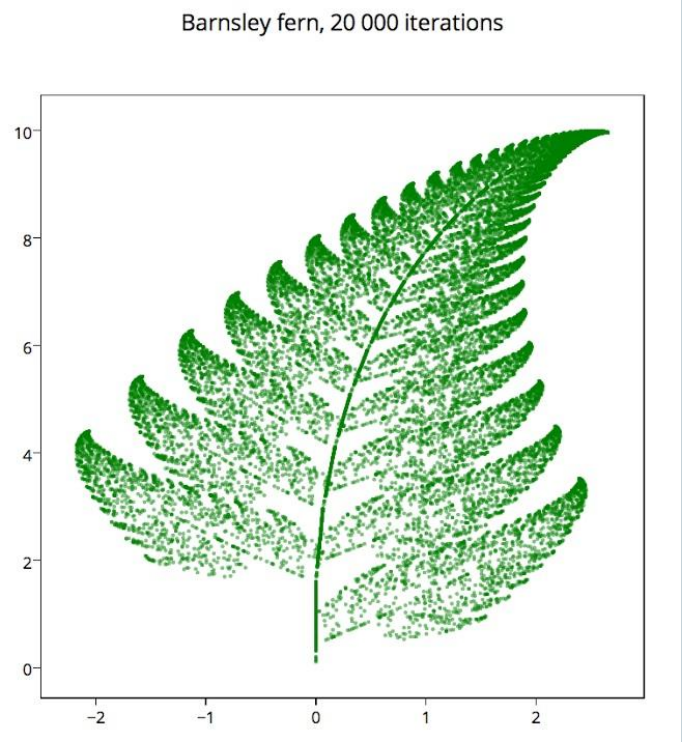
- Influx of high-dimensional data for example from medical sciences
- Desire for accurate and fast classifiers
- The challenge of integrating different types of data
- Gives more information than the methods using only one source
- A method that can efficiently work with high dimensional data regardless of data type is needed

Approach (1)

- Dynamical system is a triple (I, M, ϕ)
- I is a subset of R or Z^+ and usually tought of as time
- M is the space over wich mapping ϕ is defined (the variables are defined)
- ϕ describes the motion of a point in the space M as a function of time
- Gene expression example: $\phi : (0, T) \times R^n \rightarrow R^n$
- IFS – iterated function system (discrete dynamical system)

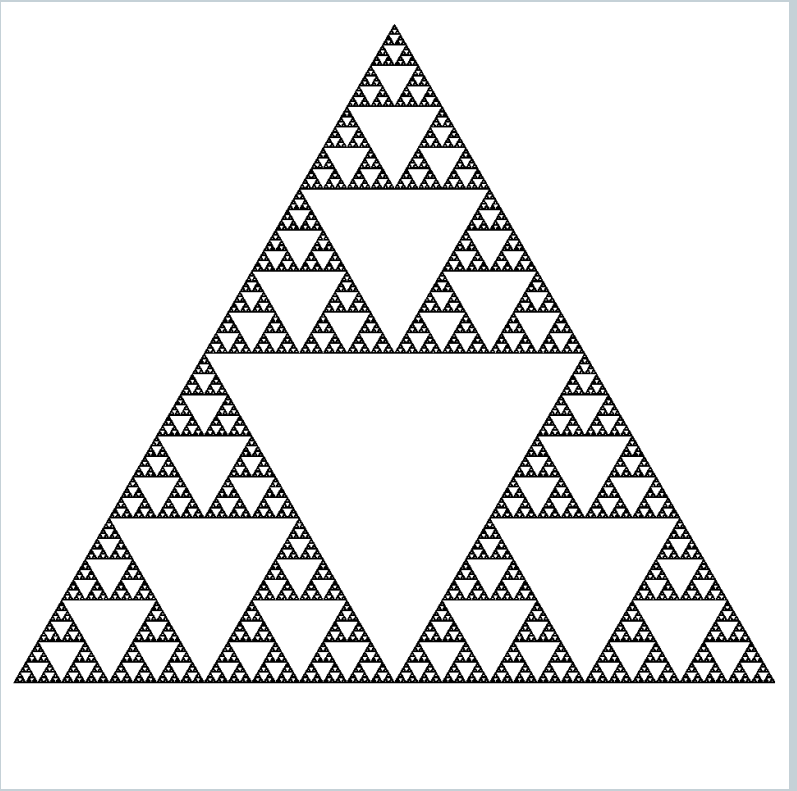
Approach (2)

- IFS contains a finite set of mappings which tranform the starting point in space M
- Each “time step” corresponds to a function and its application to the last mapping of the point
- Barnsley fern:
  $T_1(x,y)= (0.85x+0.04y, -0.04x+0.85y+1.6)$
  $T_2(x,y)=(0.2x+0.26y, 0.23x+0.22y+1.6)$
  $T_3(x,y)=(-0.15x+0.28y, 0.26x+0.24y+0.44)$
  $T_4(x,y)=(0, 0.16y)$
- Different probabilities



Approach (3)

- Chaos game (randomly iterating over functions)
- Attracting set
- Dice chaos game:
  Label vertices of a triangle:
  (12), (34), (56)
  Choose a starting vertex
  Roll the dice
  Find the midpoint between vertices
  Label it
  Roll the dice
  …
- Sierpinski triangle



Previous work

- Some dynamical systems can capture patterns through a memory-type mechanism
- Attept to study protein and DNA sequences
- Replacing the triange in chaos game with a square
- Labels A, G, C, T
- Playing the chaos game with a DNA sequence
- Detecting patterns
- Extending the idea to general datasets
- Related to functional data analysis

Data

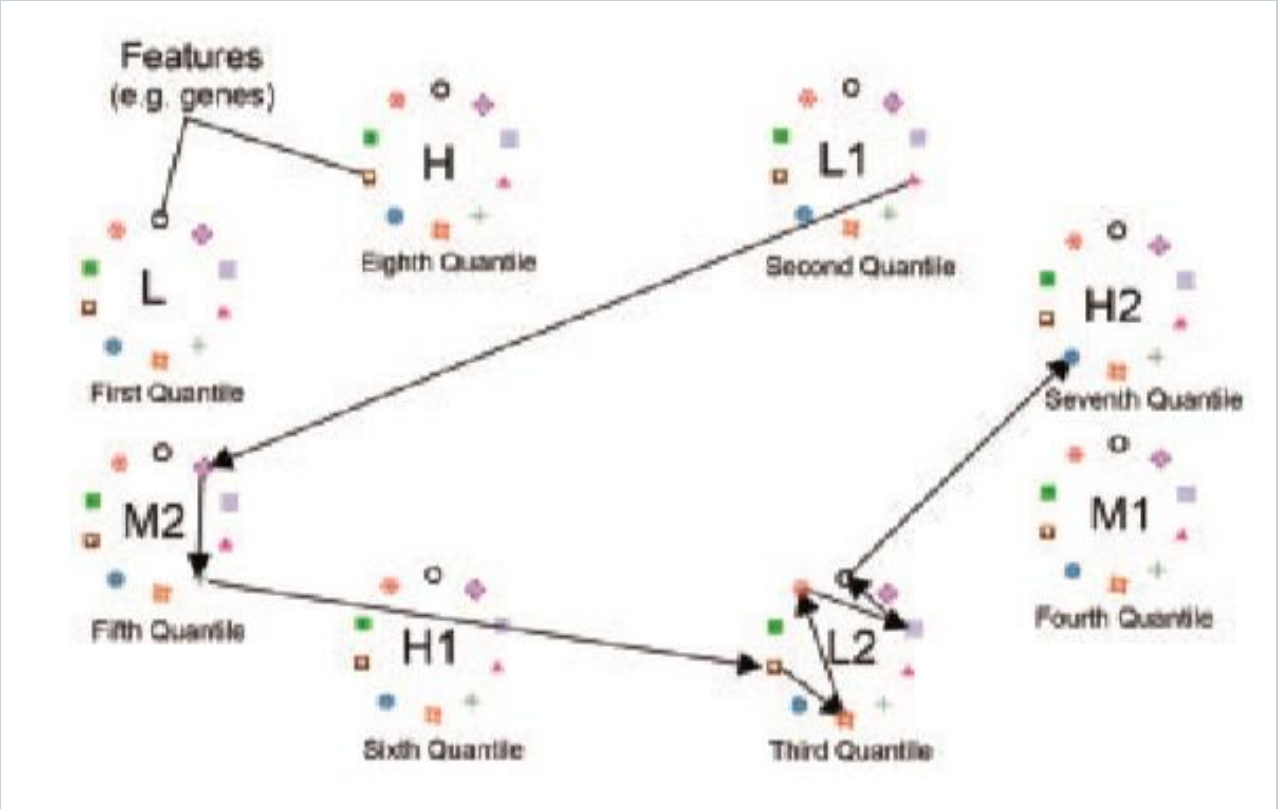
- Three datasets:
  Synthetic dataset created in Mathematica
  Gene expression lung cancer dataset (publicly available)
  Intergrated ovarian cancer dataset
- Total of 54 lung cancer patients with histological subtype adenocarcinoma (AC) and 50 with histological subtype squamos cell carcinoma (SCC)
- Ovarian dataset was gererated by their group
  14 tumors were classified as chemosensitive, 11 as chemoresistant

Algorithm (1)

- *N* subjects, *n* features
- Continuous data
- First issue is to understand how to map data to some space for chaos game
- Divide the data to eight quantiles and label {0,1,2,3,4,5,6,7} (from lowest to highest)
- Each feature can be in one of the eight states
- Alpahbet of size *n* x 8
- Choosing appropriate geometry

Algorithm (2)

- After projecting data to 2D, the clusters have to be distinguishable



Overview of the Algorithm

- Run a feature selection method
- Create a new dataset
- Discretize data into quantiles (eight is effective)
- Map each subject to a string (*f1L1.f2M2.f3M2.f4L2…)*
- Each word segment or character is given a 2D coordinate
- Run dynamical system
- Final values achieved at different feature randomization and “cut-off” features are recorded
- Final points for each subject are recorded
- Evaluating the models, the best are returned

Results on Data – Lung Cancer

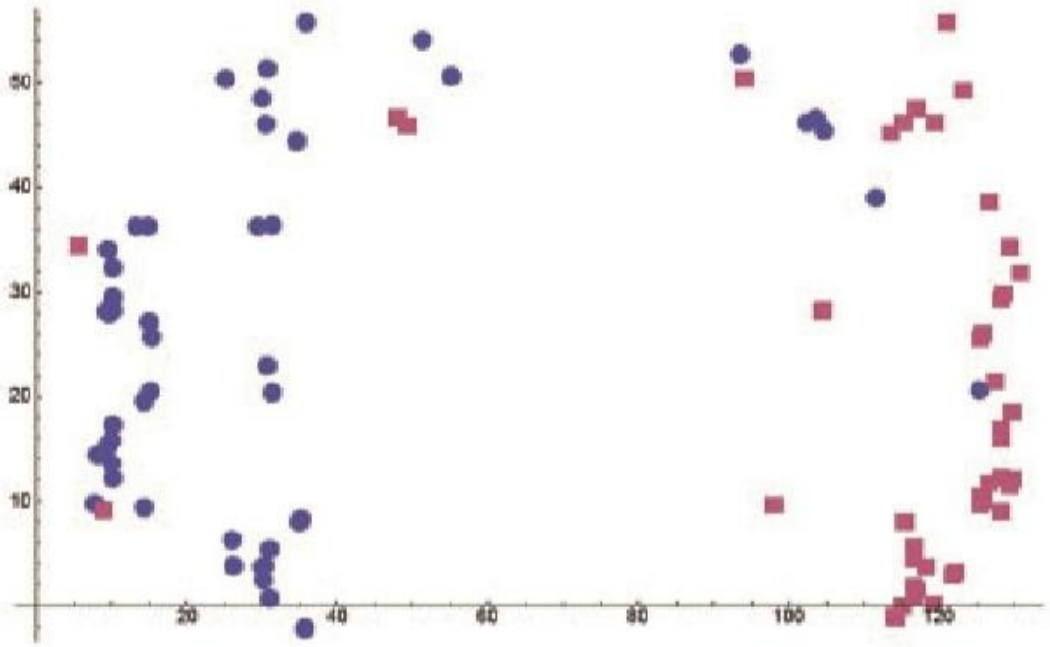


Fig. 5. The result of running Butterfly on a 77 gene lung cancer dataset

Results on Data – Lung Cancer

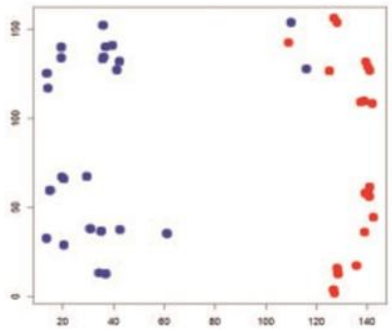


Fig. 6. A model produced on a 48 patient lung cancer dataset. Blue indicates AC cases, and red indicates SCC cases

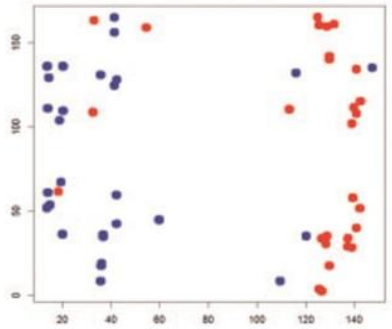


Fig. 7. The model shown in Figure 6 applied to a test set consisting of 28 AC cases and 28 SCC cases

Results on Data – Ovarian Cancer

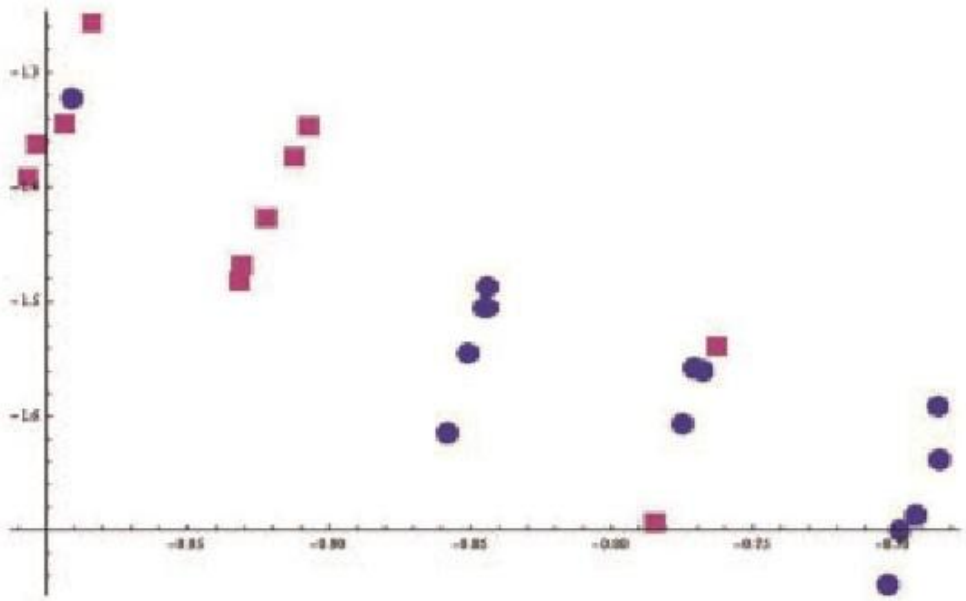


Fig. 8. Integrating mRNA, miRNA and methylation data for an ovarian cancer project. The red points indicate 11 women who responded favorably to cisplatin treatment, and the blue represents 14 women who did not

Conclusion

- An algorithm for visualization, clustering and classification of high dimensional data
- Based on simple discrete dynamical system
- Between feature selection filter and model evaluation algorithm
- Transform data into 2D representation
- Provides a human readable represenation of the relationships between subjects
- Allows both supervised and unsupervised learning

Thank you for your attention!