Fractal Dimension as a Feature Reduction Tool for Gene Expression Data

Rebecca Bernal[1], Myrine Barreiro-Areval[1], Manoj Peiris[2], Ph.D., Hansapani Rodrigo[1], Ph.D.
1. School of Mathematical and Statistical Sciences, University of Texas Rio Grade Valley, Edinburg, TX, 78539.
2. Dept of Physics and Astronomy, University of Texas Rio Grade Valley, Edinburg, TX, 78539.

INTRODUCTION

- Fractal Dimension (FD) helps to measure an object's complexity by providing a statistical index of complexity as a ratio.
- A Fractal dataset is known by its characteristics of being self similar.

Order 0 –> $(1/2)^0$
Order 1 –> $(1/2)^1$
Order 2 –> $(1/2)^2$
Order 3 –> $(1/2)^3$
Order 4 –> $(1/2)^4$
N is difference between each iteration of the fractal
r is how much smaller is each triangle in Order 1 than Order 0?

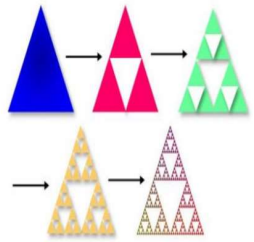

Figure 1 Fractal Dimension with Sierpinski Triangle

- The FD is relatively unaffected by redundant attributes and it can be used to detect attributes that have either linear or nonlinear correlation.

OBJECTIVE AND HYPOTHESIS

- The idea of fractal dimension can be utilized to solve dimension reduction problems associated with big data (Traina et al. 2010) .
- The objective of this study is to use fractal dimension as a dimensionality reduction tool using a combination of two scalable algorithms:
  ~ The Box Method approach and
  ~ The Fractal Dimensionality Reduction

algorithm.

- Evaluate the effectiveness of this approach as a dimensionality reduction method for gene expression data in conjunction with Random Forest Models.
- Compare its performances over Ridge Regression, LASSO Regression and Elastic Net Regression (on going work, results are yet to come).
- We hypothesized that using FD as a dimensionality reduction tool will be more effective algorithm to use compared to Ridge Regression, LASSO Regression and Elastic Net Regression.

Box-Count & Fractal Dimension Reduction Algorithms for Gene Expression Data

- For a random sample of genes with size **r** are selected from our gene expression data. Algorithm 1 was used to calculate the FD of a given data set.
- By eliminating 1 gene per time, algorithm 2 was used to calculate the partial FD for the selected sample (r times).
- By comparing FD and partial FD's, the gene with lowest difference was removed.

Algorithm 1: Box-Count Approach
Compute fractal dimension D of a dataset A

Input: normalized dataset A (N rows, with E dimensions/attributes each)

Output: fractal dimension D

Begin

- For each desirable grid-size r= 1/2j, j= 1,2, ..., l
- For each point of the data set
  - Decide which grid cell it falls in (say, the i-th cell)
  - Increment the count Ci ('occupancy')
  - Compute the sum of occupancies S(r) =∑C2i
  - Print the values of log(r) and log(S(r)) generating a plot;
  - Return the slope of the linear part of the plot as the fractal dimension D of the data set A.

End

Algorithm 2: Fractal Dimensionality Reduction Algorithm (FDR)

Input: dataset A

Output: list of attributes in the reverse order of their importance.

Begin

- 1- Compute the fractal dimension D of the whole dataset;
- 2- Initially set all attributes of the dataset as the significant ones, and the whole fractal dimension as the current D;
- 3- While there are significant attributes do:
  - 4- For every significant attribute i, compute the partial fractal dimensions pD_i using all significant attributes excluding attribute i;
  - 5- Sort the partial fractal dimensions pD_i obtained in step 4 and select the attribute a which leads to the minimum difference (current D- pD_i);
  - 6- Set the pD_i obtained removing attribute a as the current D;
  - 7- Output attribute a and remove it from the set of important attributes;

End

Data

The samples that were used for this database experiment were:

- GSE2990, Breast Cancer Gene Expression Data set that contains the details of micro array expression data for 230 breast cancer patients and 2071 genes for each patient.
- GSE44456, "Chronic high-level alcohol consumption effect on brain: post-mortem hippocampus" containing the details of microarray expression data for 39 alcoholic patients and 28870 genes.
- GSE50948, "NeOAdjuvant Herceptin (NOAH) trial: formalin-fixed, paraffin-embedded breast cancer biopsies" containing the details of microarray expression data for 156 breast cancer patients and 54675 genes.
- GSE4115, "Large airway epithelial cells from cigarette smokers with suspect lung cancer" containing the details of microarray expression data for 192 lung cancer patients and 22219 genes.

RESULTS

Breast cancer dataset (GSE2990

Accuracy

Number of Genes						
mtry	2071	1501	1251	1001	751	501
4	0.6307	0.6420	0.6307	0.6477	0.6193	0.6477
5	0.6420	0.5795	0.6534	0.642	0.642	0.642
6	0.6477	0.6591	0.6307	0.6477	0.6477	0.6477
7	0.6420	0.6420	0.6250	0.6420	0.6420	0.6420
8	0.5795	0.6477	0.6193	0.5795	0.5795	0.5795
9	0.5568	0.6136	0.6705	0.5568	0.5568	0.5568
11	0.6307	0.5739	0.6080	0.307	0.307	0.3070
12	0.6193	0.5909	0.6136	0.6193	0.6193	0.6193
13	0.6023	0.5909	0.6420	0.6023	0.6023	0.6193
14	0.642	0.5966	0.5909	0.6420	0.6420	0.6420
20	0.5852	0.6136	0.5568	0.5852	0.5852	0.5852

mtry: Number of variables available for splitting at each tree node.

Sensitivity

Number of Genes						
mtry	2071	1501	1251	1001	751	501
4	0.4546	0.5238	0.4286	0.5455	0.3333	0.5333
5	0.5294	0.3684	0.5789	0.5294	0.5294	0.5294
6	0.5625	0.6667	0.4667	0.5625	0.5625	0.5625
7	0.5385	0.5185	0.4546	0.5385	0.5385	0.5385
8	0.3958	0.5625	0.4516	0.3958	0.3958	0.3958
9	0.3833	0.3750	0.6875	0.3833	0.3833	0.3833
11	0.4894	0.3590	0.4324	0.4894	0.4894	0.4894
12	0.4546	0.4167	0.4474	0.4546	0.4546	0.4546
13	0.4531	0.4091	0.5789	0.5294	0.4531	0.4531
14	0.5333	0.2941	0.3889	0.5333	0.5333	0.5333
20	0.4118	0.4375	0.2941	0.4118	0.4118	0.4118

mtry: Number of variables available for splitting at each tree node.

Specificity

Number of Genes						
mtry	2071	1501	1251	1001	751	501
4	0.6424	0.6581	0.4639	0.6623	0.6347	0.6712
5	0.6541	0.6377	0.6624	0.6541	0.6541	0.6541
6	0.6562	0.6585	0.6460	0.6563	0.6563	0.6563
7	0.6503	0.6644	0.6494	0.6503	0.6503	0.6503
8	0.6484	0.65625	0.6552	0.6484	0.6484	0.6484
9	0.6466	0.6375	0.6688	0.6466	0.6466	0.6466
11	0.6822	0.6350	0.6547	0.6822	0.6822	0.6822
12	0.6573	0.6562	0.6594	0.6573	0.6573	0.6573
13	0.6875	0.6515	0.6788	0.6875	0.6875	0.6875
14	0.6522	0.6290	0.6429	0.6522	0.6522	0.6522
20	0.6560	0.6528	0.6197	0.6560	0.6560	0.6560

mtry: Number of variables available for splitting at each tree node.

Alcohol consumption dataset (GSE44456)

(A)

mtry	# of Genes		
	22269	16599	8706
4	0.5556	0.6667	0.6667
6	0.5556	0.5556	0.5556
8	0.5556	0.6667	0.6667
11	0.5556	0.6667	0.5556
13	0.6667	0.6667	0.6667
20	0.6667	0.6667	0.6667

(B)

mtry	# of Genes		
	22269	16599	8706
4	0.6000	0.6667	0.6667
6	0.6000	0.6000	0.6000
8	0.6000	0.6667	0.6667
11	0.6006	0.6667	0.6000
13	0.6667	0.6667	0.6667
20	0.6667	0.6667	0.6667

(C)

mtry	# of Genes		
	22269	16599	8706
4	0.5000	0.6667	0.6667
6	0.5000	0.5000	0.5000
8	0.5000	0.6667	0.6667
11	0.5000	0.6667	0.5000
13	0.6667	0.6667	0.6667
20	0.6667	0.6667	0.6667

Lung cancer dataset (GSE4115)

(A)

mtry	# of Genes		
	22215	20915	18415
4	0.5957	0.5957	0.5957
6	0.5745	0.5957	0.6383
8	0.617	0.5957	0.6383
11	0.5745	0.5745	0.617
13	0.5745	0.5957	0.5745
20	0.5957	0.5957	0.5957

(B)

mtry	# of Genes		
	22215	20915	18415
4	0.5862	0.5862	0.6154
6	0.5667	0.6000	0.6061
8	0.6000	0.6154	0.6207
11	0.5667	0.5714	0.6000
13	0.5667	0.5862	0.5862
20	0.6071	0.6000	0.5806

(C)

mtry	# of Genes		
	22215	20915	18415
4	0.6111	0.6111	0.6190
6	0.5882	0.6471	0.7143
8	0.6471	0.6190	0.6667
11	0.5882	0.5789	0.6471
13	0.5882	0.6111	0.6111
20	0.6316	0.6471	0.6250

Lung cancer dataset (GSE50948)

(A)

mtry	# of Genes		
	54675	52675	50675
4	0.5495	0.5313	0.5135
6	0.4234	0.5313	0.4504
8	0.5585	0.4414	0.4864
11	0.5405	0.4777	0.5045
13	0.4954	0.4954	0.4324
20	0.5585	0.4594	0.5045

(B)

mtry	# of Genes		
	54675	52675	50675
4	0.7333	0.80	0.6666
6	0.6666	0.70	0.6333
8	0.6666	0.70	0.6333
11	0.5333	0.7333	0.6666
13	0.5333	0.5333	0.6333
20	0.6	0.6	0.7

(C)

mtry	# of Genes		
	54675	52675	50675
4	0.2000	0.0660	0.2000
6	0.2000	0.2500	0.066
8	0.2000	0.4000	0.2666
11	0.1330	0.3333	0.2666
13	0.2666	0.2666	0.2000
20	0.3333	0.2666	0.2000

Figure 2. Random Forest Evaluations for four gene expression data sets **.** **(A)** Accuracy = (TN + TP)/(TN+TP+FN+FP), **(B)** Sensitivity = TP/(TP + FN), **(C)** Specificity = TN/(TN + FP) , values for Full data (second column in each table) and FD reduced gene expression data (third and fourth columns in each table). "mtry" represents the number of variables available for splitting at each tree node in each RF model.

DISCUSSION

- We have explored the FD based dimension reduction technique for gene expression data for the first time.
- This take into account the correlation between the genes and hence will help to remove redundant genes.
- For the most part, this techniques resulted in a same or at least slightly better improvement in accuracy, sensitivity and specificity values for gene expression data.
- This need to be further evaluated for different gene reduction values (currently, we have tried only two different values).

BIBLIOGRAPHY

- C. Traina, A. Traina, L. Wu, and C. Faloutsos, "***Fast feature selection using the fractal dimension***". In Proc. of XV Brazilian Symposium on Databases, 2000
- Robert L. Devaney, "***Fractal Dimension***", Sun Apr 2 14:31:18 EDT 1995
- Wang Y, Klijn JG, Zhang Y, Sieuwerts AM et al., "***Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer***", Lancet 365(9460) 2005.
- National Center for Biotechnology Information | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/