Unsupervised Feature Selection Based on the Morisita Index

Jean Golay and Mikhail Kanevski

Institute of Earth Surface Dynamics (IDYST)
Faculty of Geosciences and Environment (FGSE)
University of Lausanne, Switzerland

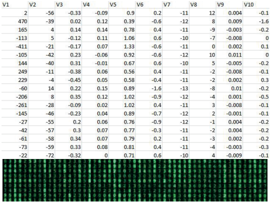
EGU General Assembly 2016

Introduction

High-Dimensional Data Sets

←Variables / Features→

Instances



V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
2	-56	-0.33	-0.09	0.9	0.2	-11	12	0.004	-0.1
470	-39	0.02	0.12	0.39	-0.6	-12	8	0.009	-1.6
165	4	0.14	0.14	0.78	0.4	-11	-9	-0.003	-0.2
-113	5	-0.12	0.11	1.06	0.6	-10	-7	-0.008	0
-411	-21	-0.17	0.07	1.33	-0.6	-11	0	0.002	0.1
-105	-42	0.23	-0.06	0.92	-0.6	-12	10	0.011	0
144	-40	0.31	-0.01	0.67	0.6	-10	5	-0.005	-0.2
249	-11	-0.38	0.06	0.56	0.4	-11	-2	-0.008	-0.1
229	-4	-0.45	0.05	0.58	-0.4	-11	-2	0.002	0.3
-60	14	0.22	0.15	0.89	-1.6	-13	-8	0.01	-0.2
-206	8	0.35	0.12	1.02	-0.9	-12	-4	0.001	-0.5
-261	-28	-0.09	0.02	1.02	0.4	-11	3	-0.008	-0.1
-145	-46	-0.23	0.04	0.89	-0.7	-12	2	-0.001	-0.1
-27	-55	0.2	0.06	0.76	-0.9	-12	-1	0.004	-0.2
-42	-57	0.3	0.07	0.77	-0.3	-11	-2	0.004	-0.2
-61	-58	0.34	0.07	0.79	0.2	-11	-3	0.002	-0.2
-73	-59	0.33	0.08	0.81	0.4	-11	-4	-0.003	-0.3
-22	-72	-0.32	0	0.71	0.6	-10	4	-0.009	-0.1

Issues

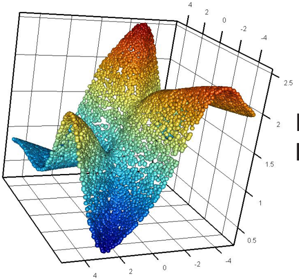
1. Curse of dimensionality
2. Computer performance
3. Data visualization
4. Interpretability of the results

Solutions

1. PCA
2. MDS
3. etc.

A New Solution

1. The concept of Intrinsic Dimension (ID)
2. The Morisita estimator of ID
3. An ID-based algorithm for selecting the smallest subset of features conveying all the information content of a data set



ID=2
E=3

Outline

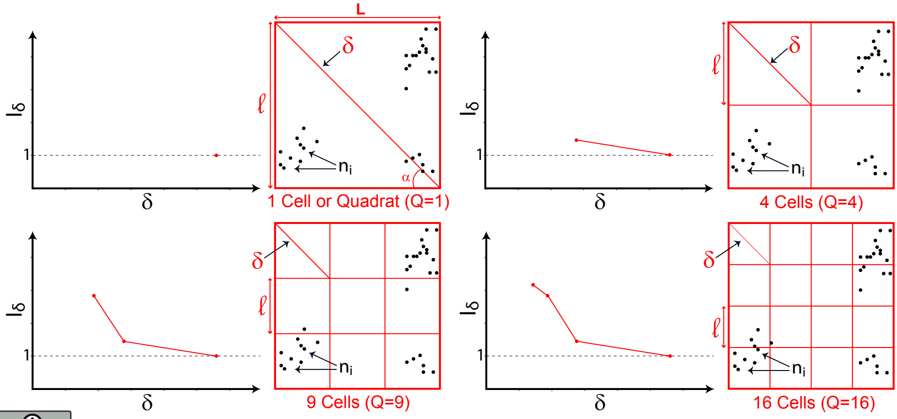
1. The Morisita index
   - Calculation
2. The Morisita Estimator of ID
   - Calculation
3. ID-based Feature Selection
   - ID and Redundancy
   - The Proposed Algorithm
   - A Simulated Case Study
   - Real Case Studies
4. Conclusion

Outline

1. The Morisita index
2. The Morisita Estimator of ID
3. ID-based Feature Selection
4. Conclusion

Calculation

$$I_\delta = Q \frac{\sum_{i=1}^{Q} n_i(n_i-1)}{N(N-1)}$$



1 Cell or Quadrat (Q=1)

4 Cells (Q=4)

9 Cells (Q=9)

16 Cells (Q=16)

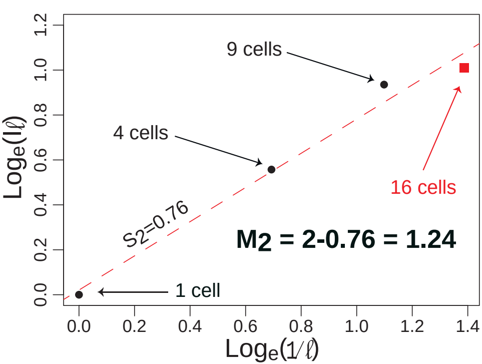
Outline

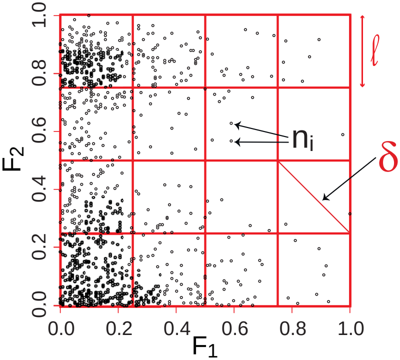
1. The Morisita index
2. The Morisita Estimator of ID
3. ID-based Feature Selection
4. Conclusion

Calculation

$$M_2 = E - S_2$$

The concept of ID is extended to non-integer (fractal) dimensions.





Outline

1 The Morisita index

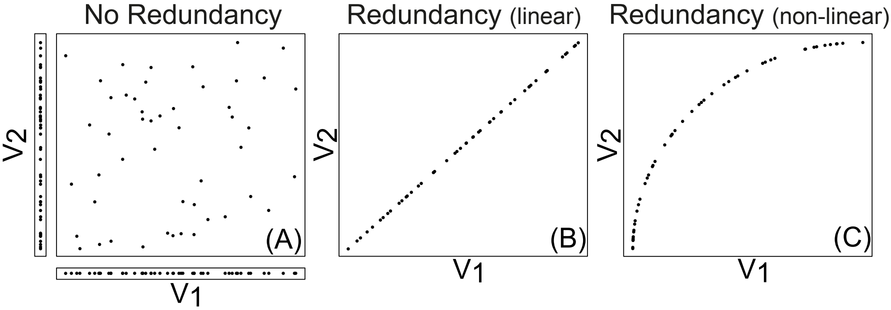
2 The Morisita Estimator of ID

3 ID-based Feature Selection

4 Conclusion

ID and Redundancy

V_1 and V_2 are two uniformly distributed variables and one has that:
$ID(V_1, V_2) \approx ID(V_1) + ID(V_2) \approx 1 + 1 = 2$ (see (A))
$ID(V_1, V_2) \approx ID(V_1) \approx ID(V_2) \approx 1$ (see (B) and (C))

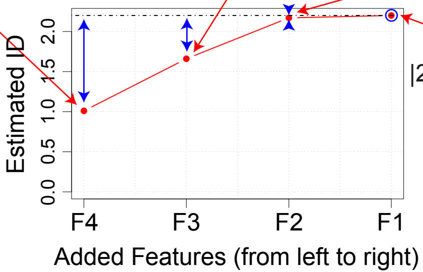


Redundant features (variables) do not contribute to the data ID
Idea (Traina's work): select the features which increase the data ID.

The Proposed Algorithm

A = {F1,F2,F3,F4} and M_2(A) = 2.20

Step 1	Step 2	Step 3
|2.20 - M_2(F1)| = 1.20	|2.20 - M_2(F4,F1)| = 1.14	|2.20 - M_2(F4,F3,F1)| = 0.52
|2.20 - M_2(F2)| = 1.34	|2.20 - M_2(F4,F2)| = 0.59	|2.20 - M_2(F4,F3,F2)| = 0.03
|2.20 - M_2(F3)| = 1.30	|2.20 - M_2(F4,F3)| = 0.54	
|2.20 - M_2(F4)| = 1.19		



Step 4

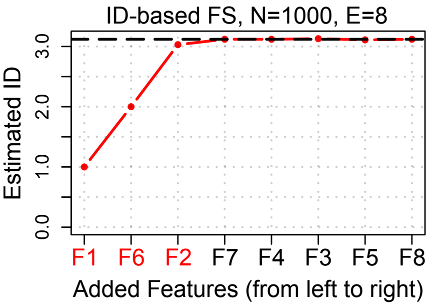
|2.20 - M_2(F4,F3,F2,F1)| = 0.00

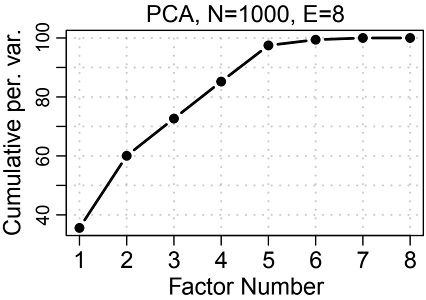
F1 is redundant, since it hardly contributes to increasing the data ID

A Simulated Case Study: the Input Space of the Butterfly Data Set

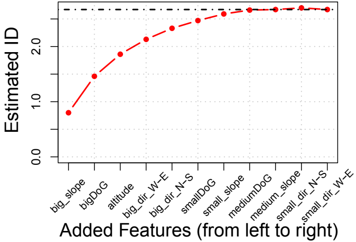
$F_1, F_2, F_6 \in]-5, 5[$ $\quad F_3 = \log_{10}(F_1 + 5)$ $\quad F_4 = F_1^2 - F_2^2$

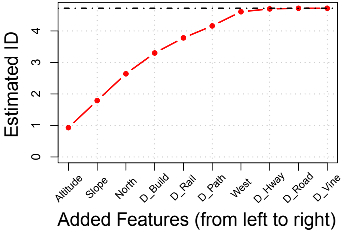
$F_5 = F_1^4 - F_2^4$ $\quad F_7 = \log_{10}(F_6 + 5)$ $\quad F_8 = F_6 + F_7$





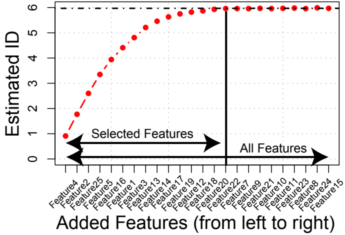
Real Case Studies I


Windspeed: N=127, E=11


Forest Fires: N=2275, E=10

Permafrost: N=54674, E=24

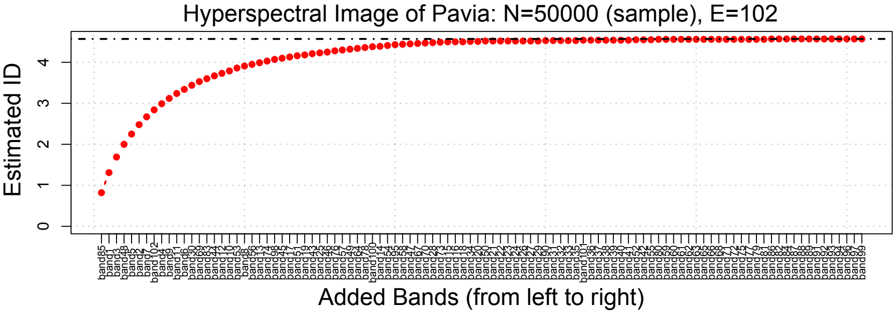
The data set was kindly provided by Nicola Deluigi



The information content was assessed by means of random forest

	All Features	Selected Features
Error rate (%)	14.93 (3.13)	14.40 (2.63)

Real Case Studies II



Outline

1. The Morisita index
2. The Morisita Estimator of ID
3. ID-based Feature Selection
4. Conclusion

Conclusion

Take-Home Message

The concept of Intrinsic Dimension (ID) can help find solutions to the issues raised by large data sets.

- C. Traina Jr., A. J. M. Traina, L. Wu, C. Faloutsos, Fast feature selection using fractal dimension, *Proceedings of the XV Brazilian Symposium on Databases (SBBD)*, pp. 158-171, 2000.
- J. Golay, M. Leuenberger, M. Kanevski, Feature Selection for Regression Problems Based on the Morisita Estimator of Intrinsic Dimension, *arXiv:1602.00216*, 2016.
- J. Golay and M. Kanevski, A new estimator of intrinsic dimension based on the multipoint Morisita index, *Pattern Recognition*, 48(12):4070-4081, 2015.
- J. Golay, M. Kanevski, C. D. Vega Orozco and M. Leuenberger, The multipoint Morisita index for the analysis of spatial patterns, *Physica A*, 406:191-202, 2014.