 doi:10.20944/preprints202102.0260.v2

SimulaD: A Novel Feature Selection Heuristics For Discrete Data

Mohammad Reza Besharati

Mohammad Izadi

1- besharati@ce.sharif.edu, PhD Candidate, Sharif University of Technology, Tehran, Iran, Corresponding Author.
2- izadi@sharif.edu, Assistant Professor, Sharif University of Technology, Tehran, Iran.

Abstract

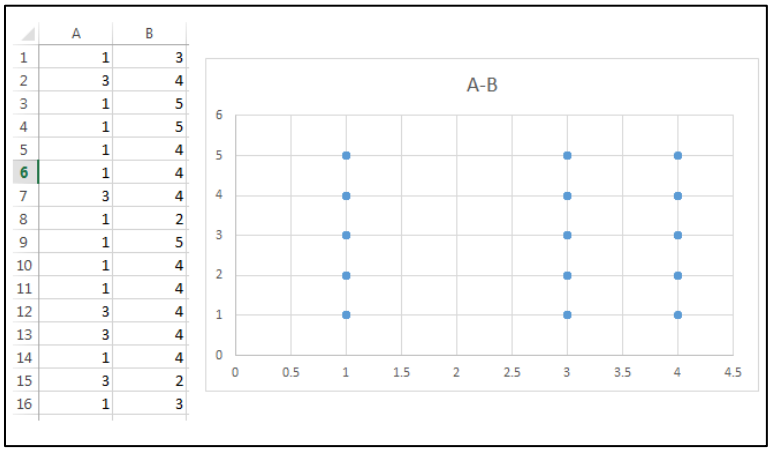
By applying a running average (with a window-size= d), we could transform Discrete data to broad-range, Continuous values. When we have more than 2 columns and one of them is containing data about the tags of classification (Class Column), we could compare and sort the features (Non-class Columns) based on the R2 coefficient of the regression for running averages. The parameters tuning could help us to select the best features (the non-class columns which have the best correlation with the Class Column). “Window size” and “Ordering” could be tuned to achieve the goal. this optimization problem is hard and we need an Algorithm (or Heuristics) for simplifying this tuning. We demonstrate a novel heuristics, Called Simulated Distillation (SimulaD), which could help us to gain a somehow good results with this optimization problem.

Keywords: Feature Selection, Discrete Data, Heuristics, Running average

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 March 2021 doi:10.20944/preprints202102.0260.v2

Introduction

Suppose that we have 2 column of discrete data (Column A and Column B) about software quality from users' perspective with Likert-scale values [1] [2]. We wish to find the probable correlation between these two columns. If the range of discrete values is limited, then we couldn't shape a sufficient space of locus points to run regression algorithms (see figure-1).



	A	B
1	1	3
2	3	4
3	1	5
4	1	5
5	1	4
6	1	4
7	3	4
8	1	2
9	1	5
10	1	4
11	1	4
12	3	4
13	3	4
14	1	4
15	3	2
16	1	3

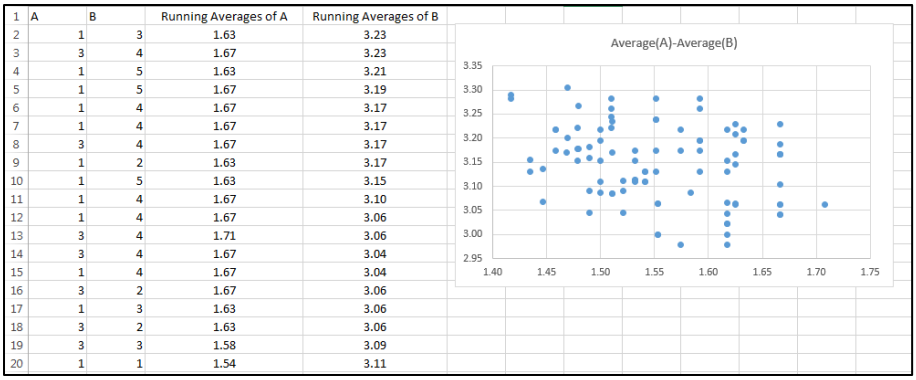
Figure 1 - the Locus space for limited-range, discrete values of Columns A and B.

We need a schema to map these limited-range, discrete values to broad-range, continuous values. This schema must conserve the necessary characteristics of initial values to show us any probable correlation between values.

The Proposed Method

By applying a running average (with a window-size= d), we could transform the data to broad-range, Continuous values (see figure-2). It's could be considered as a type of continuous measuring of discrete data. Then we could apply regression algorithms to investigate the inherent correlation between these two sets of values (see figure-3). A real-world example is provided in figure-4. We could consider each point of the resulting continuous space locus, as a representation of a micro-community with d users population.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 March 2021 doi:10.20944/preprints202102.0260.v2



	A	B	Running Averages of A	Running Averages of B
2	1	3	1.63	3.23
3	3	4	1.67	3.23
4	1	5	1.63	3.21
5	1	5	1.67	3.19
6	1	4	1.67	3.17
7	1	4	1.67	3.17
8	3	4	1.67	3.17
9	1	2	1.63	3.17
10	1	5	1.63	3.15
11	1	4	1.67	3.10
12	1	4	1.67	3.06
13	3	4	1.71	3.06
14	3	4	1.67	3.04
15	1	4	1.67	3.04
16	3	2	1.67	3.06
17	1	3	1.63	3.06
18	3	2	1.63	3.06
19	3	3	1.58	3.09
20	1	1	1.54	3.11

Figure2 -the Locus space for broad-range, continuous values of running averages with window-size=50.

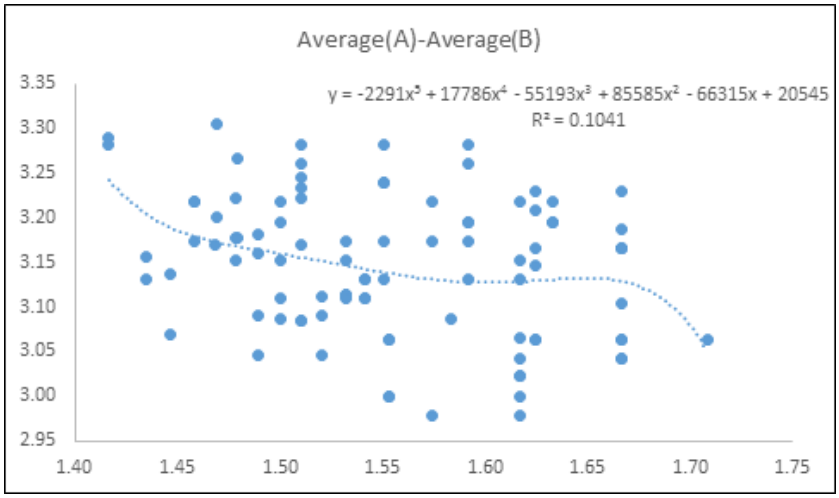


Figure 3- The Regression was applied to investigate the correlation.

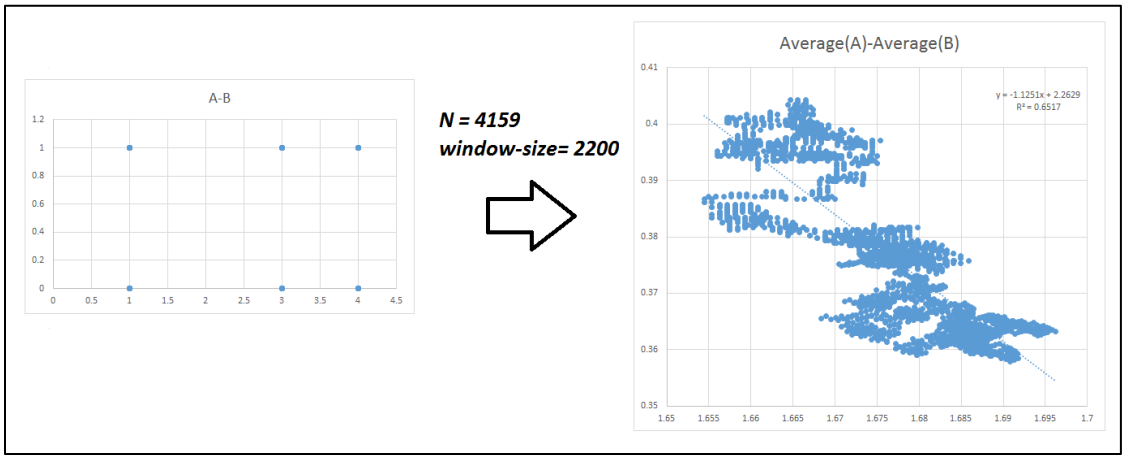


Figure 4- from discrete values to concurrent regression

By varying the window size (d), the regression factor R^2 is varying. For different datasets, we could plot different d-R^2 diagrams. Extremum points of these plots are depicting an inherent characteristics feature of the dataset (see figure-5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 March 2021 doi:10.20944/preprints202102.0260.v2

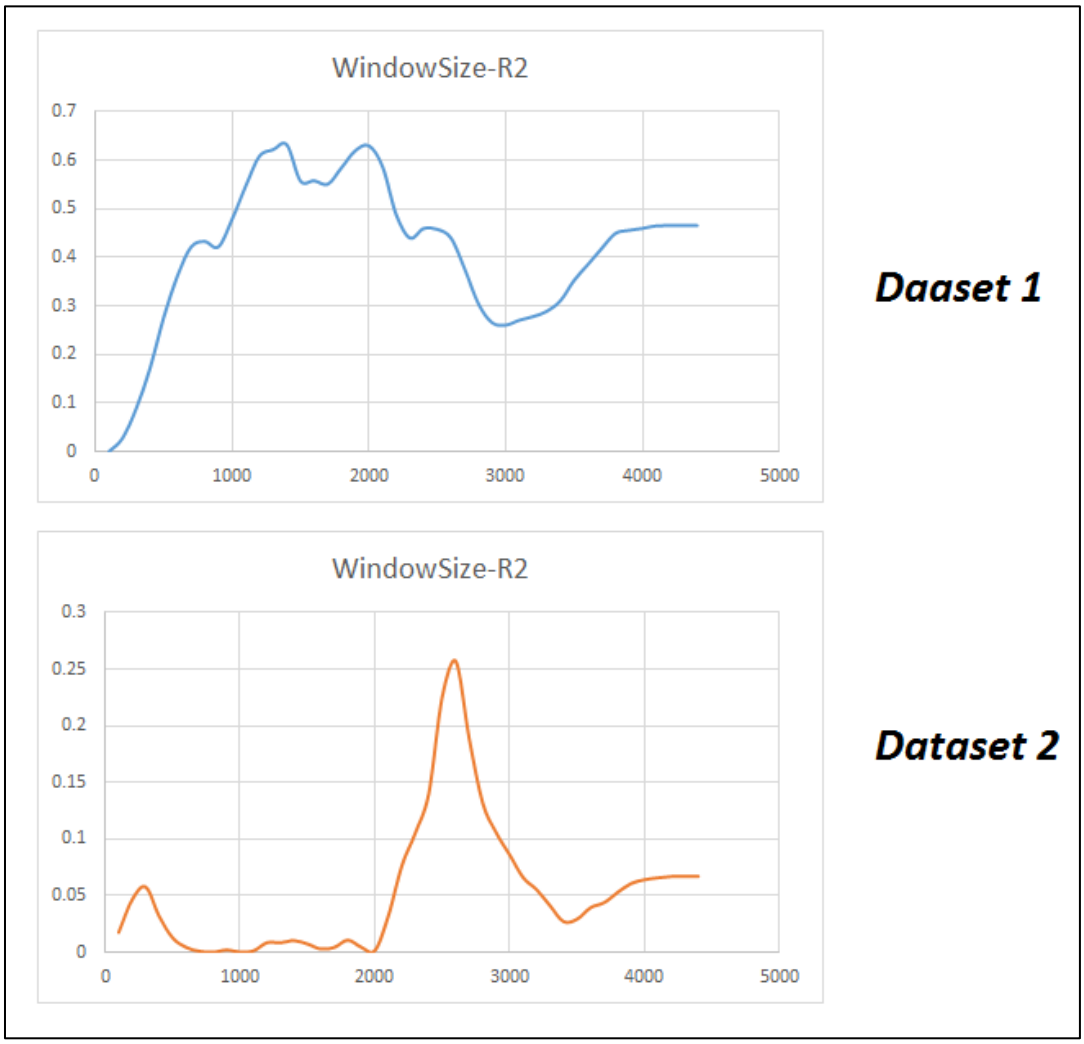


Figure 5- WindowSize-R2 diagram, plotted for two different datasets.

Comparing with Random base-line

We could examine the level of correlation by comparing the R2 coeficient for two different settings: 1) when columns are filled with the running averages of under study data, 2) the columns are filled with running averages of a randomly-generated base-line data.

Ordering of Data Effects the Correlation Results

Each ordering of the data yields a different R2 coefficient. So after the window size, the ordering is another parameter for tuning the results. An appropriate ordering could show us the inherent correlation between two columns (albeit after applying the regression). A set of N data items has N! different orderings and we couldn't check each one. So we need

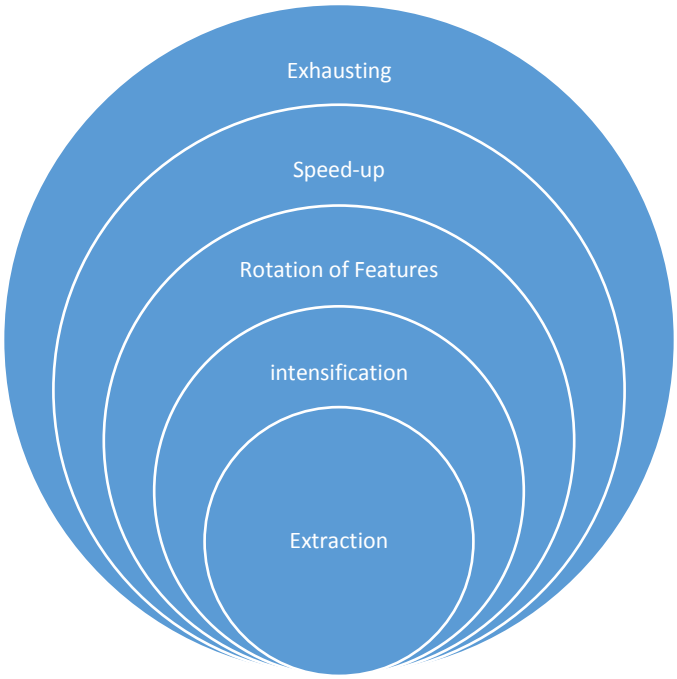
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 March 2021 doi:10.20944/preprints202102.0260.v2

some algorithm (or Heuristics) to select appropriate ordering of data. As a simple one, we could average (or select the optimum from) the results of some random-selected samples of the "Ordering Space" of the data.

Feature Selection Method

When we have more than 2 columns and one of them is containing data about the tags of classification (Class Column), we could compare and sort the features (Non-class Columns) based on the R2 coefficient of the regression for running averages.

The parameters tuning could help us to select the best features (the non-class columns which have the best correlation with the Class Column). "Window size" and "Ordering" could be tuned to achieve the goal. Again our optimization problem is hard and we need an Algorithm (or Heuristics) for simplifying this tuning. We demonstrate a novel heuristics, Called Simulated Distillation (SimulaD), which could help us to gain a somehow good results with this optimization problem.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 March 2021 doi:10.20944/preprints202102.0260.v2

```
for (int o = 0; o < 200; o++) {// 1) Exhausting

      LinearRegression.maxR2=0;

      for (int z = 1; z < 10; z++) { // 2) Speed-Up

            for (int u = 1; u < 84; u++) {// 3) Rotation of Features

                  for (int y = 0; y < 3; y++) {// 4) Intensification

                        // 5) Extraction:
                        //Randomize order and Calculate R2
                        //for Feature u with WindowSize=400
                        computeNewOrderAndRegression(u, 400, DataStore);

                              }

                        }

                  }
            }
```

Definition 1. Winner(E) = F <=>

F has the best R2 coefficient among features in Exhausting Epoch E.

Definition 2.

$$\text{Win-ratio (F)} = \frac{\#\text{ of Exhausting } Epochs\ where\ F\ is\ winner}{Total\ Number\ of\ Exhausting\ Epochs}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 March 2021 doi:10.20944/preprints202102.0260.v2

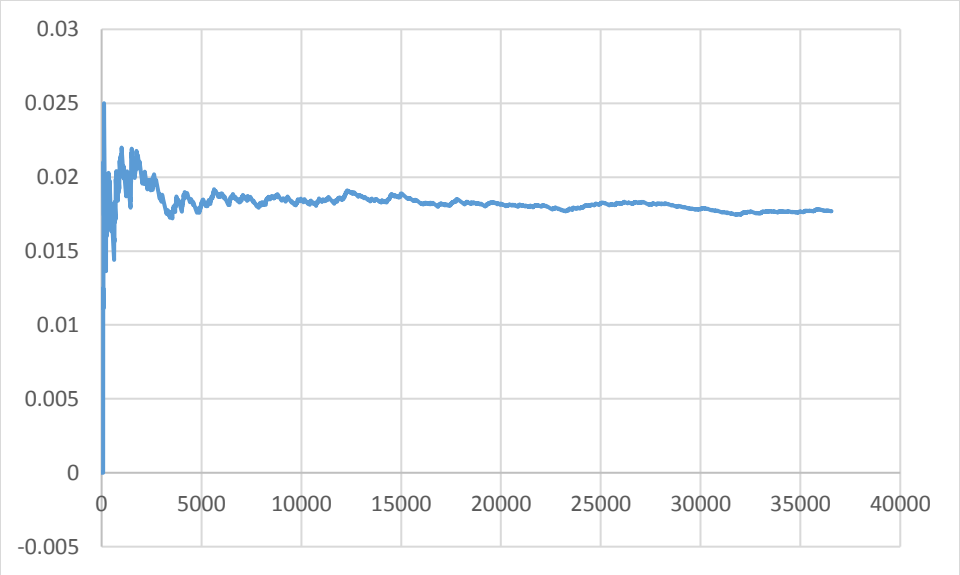


Figure 6- Win ratio in the Exhausting Epochs, depicted for one of the features.

Conclusion

By applying a running average (with a window-size= d), we could transform the data to broad-range, Continuous values. It's could be considered as a type of continuous measuring of discrete data. We could compare and sort the features (Non-class Columns) based on the R2 coefficient of the regression for running averages. We have demonstrated a novel heuristics, Called Simulated Distillation (SimulaD), which could help us to gain a somehow good results with optimization problem of "Window Size" and "Ordering".

References

[1] Besharati, Mohammad Reza, and Mohammad Izadi. "KARB Solution: Compliance to Quality by Rule Based Benchmarking." *arXiv preprint arXiv:2007.05874* (2020).

[2] R. Likert, A technique for the measurement of attitudes, Arch. Psychol. (1932).