 doi:10.20944/preprints202102.0260.v3

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, Associate Professor, Sharif University of Technology, Tehran, Iran.

Abstract

For discrete big data which have a limited range of values, Conventional machine learning methods cannot be applied because we see clutter and overlapping of classes in such data: many data points from different classes overlap. In this paper we introduce a solution for this problem through a novel heuristics method. 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, Quantification of Qualities.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

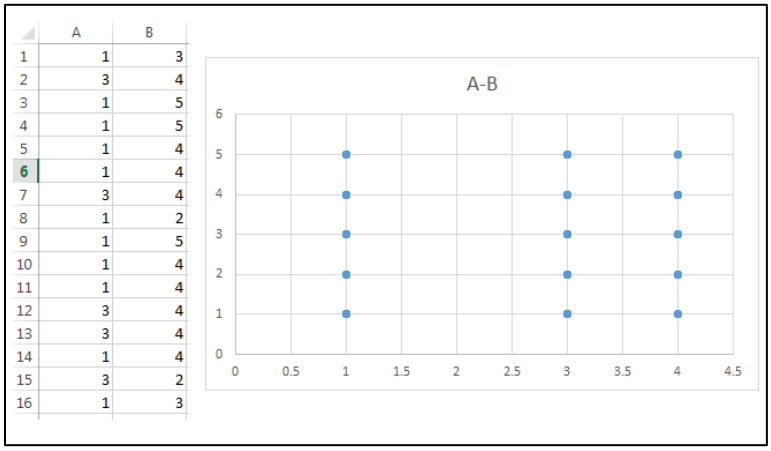
Introduction

There are numerous previous heuristics for feature selection methods [1]. Heuristics are based on human intuitions for solving technical problems [2]. Here we provide a new and novel heuristic for Feature Selection Method.

For discrete big data which have a limited range of values, Conventional machine learning methods cannot be applied because we see clutter and overlapping of classes in such data: many data points from different classes overlap. In this paper, by applying a novel heuristics for feature selection method, we overcome this problem. We use the moving average filter [3] and linear regression [4] to achieve the goal.

The problem description

Suppose that we have 2 column of discrete data (Column A and Column B) about software quality from users' perspective with Likert-scale values [5] [6] [7]. 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 [8] (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-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

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.

	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

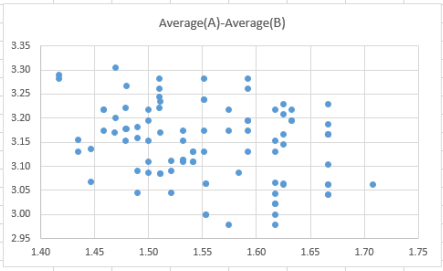


Figure2 -the Locus space for broad-range, continuous values of running averages with window-size=50.

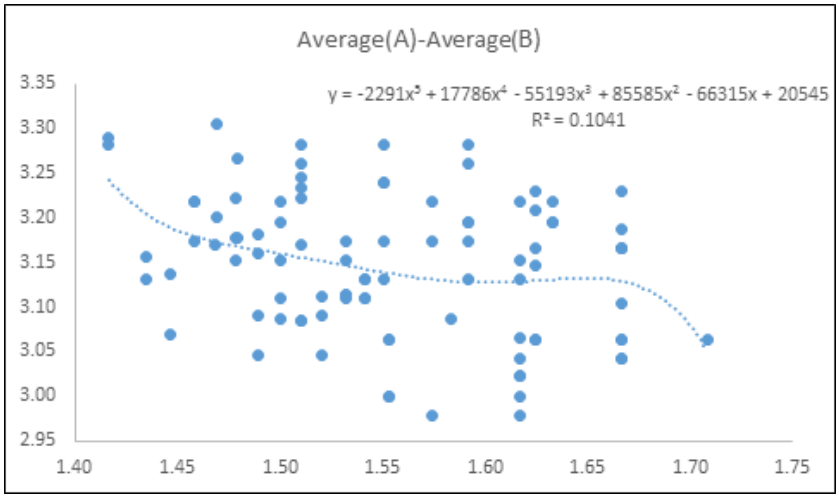


Figure 3- The Regression was applied to investigate the correlation.

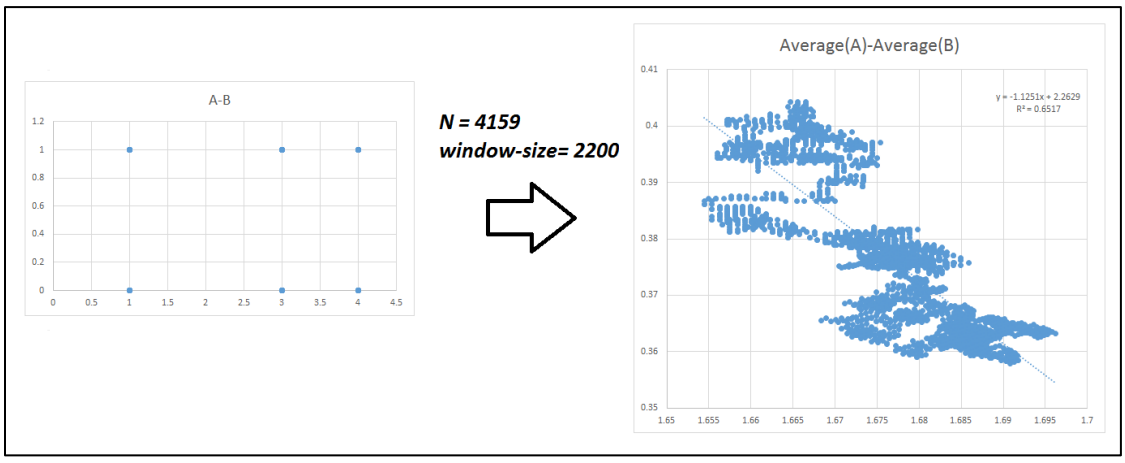


Figure 4- from discrete values to concurrent regression

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

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).

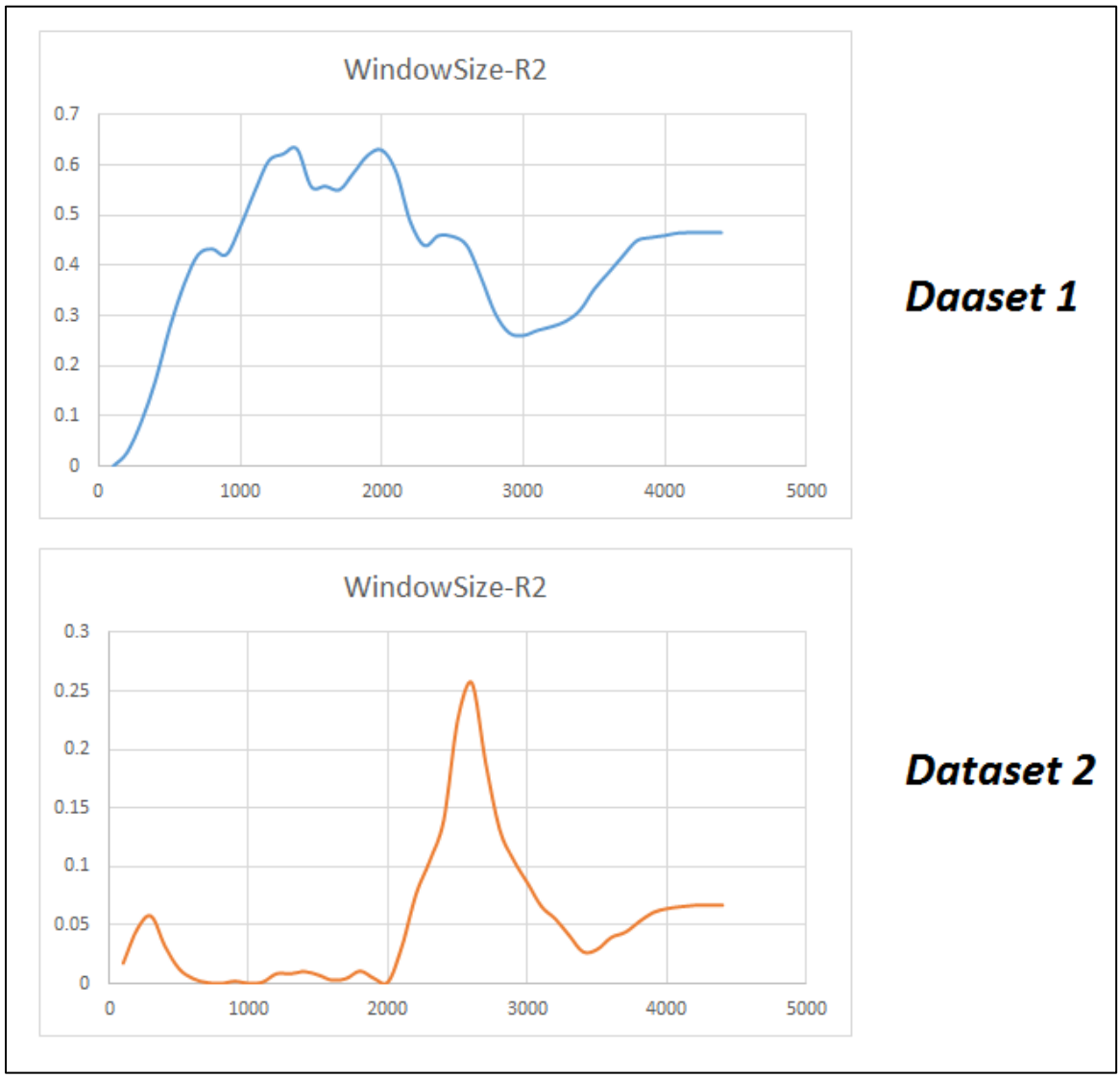


Figure 5- WindowSize-R2 diagram, plotted for two different datasets.

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

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 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.

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

      LinearRegression.maxR2=0;

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

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

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

                        // Extraction
                        computeNewOrderAndRegression(u, 400, makhzan);

                              }

                        }

                  }
            }
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

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

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

Definition 2.

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

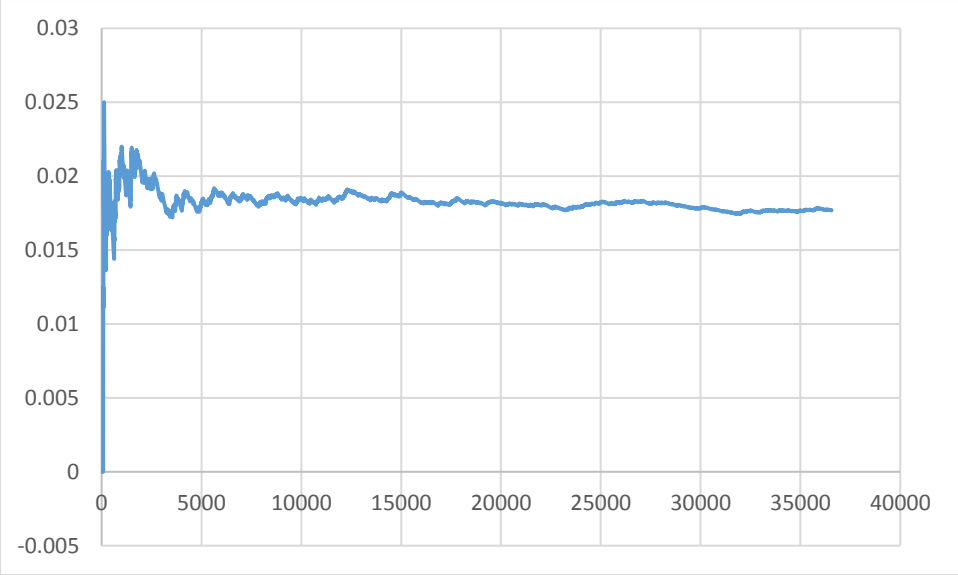


Figure 6- Win ratio in the Exhausting Epochs, depicted for one of the features.

We used Iranian National Computing Grid services[1] (80 computing cores) to do the computations.

Algorithm Steps

1- Based on intuitions, consider a window size d for data.
2- For k_1 times: `// Exhausting`
   a. MaxR2=0;
   b. Winner=null;
   c. For k_2 times: `// Speed-Up`
      i. For each non-class column u (feature u): `// Rotation`
         1. For k_3 times: `// Intensification and Extraction`
            a. Randomly reorder the data records.
            b. Calculate running averages (with window size=d) for non-class column u (feature u).
            c. Compute R2 for regression on u-running-averages and the class column.
            d. if MaxR2 < resulting_R2_from_previous_step:
               i. MaxR2= resulting R2; Winner=u.
   d. Print MaxR2 and Winner.

[1] https://turin.ipm.ir/en/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

Computational Complexity

For a dataset with N features and M records, the computational complexity of proposed heuristics algorithm is: O ($k_1k_2k_3$NM).

Discussion on Application domain

When we want to study a number of quality variables for a system, we can use different methods of quality quantification: Likert scale quality questionnaires [9], fuzzy logic [10] [11], totaling over system segments, statistical distribution approximation [12], Continuous signal approximation from discrete samples [13] and so on.

Our proposed method, which is suitable for many systems, especially complex socio-technical systems, is as follows: Using Likert-scale quality spectrum questionnaires to collect quantitatively discrete data about system quality variables, then convert this Discrete data to continuous data that are suitable for machine learning (by applying SimulaD algorithm on them).

The Advantage over other methods

Our proposed algorithm uses the concept of micro-community [5]. Each micro-community is like a field of probability around each of the records. So in fact our approach uses the mathematical concept of "probability field", albeit conceptually rather than technically, instead of using other mathematical objects (such as sets in fuzzy logic, signals in signal processing, totaling in statistics and so on). The stochastic nature of probability fields are very match with “quality fluctuations” in the real world. So it could be a suitable quantitative model for describing quality variables of complex systems.

Focusing on micro-communities could unravel hidden topological structures in data (for an example, see figure-7). So it could be useful in the topological data analysis.

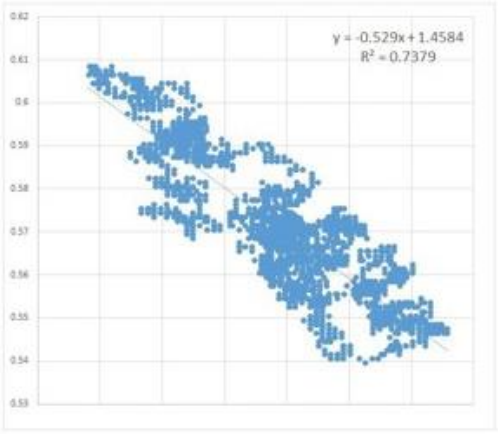


Figure 7- depicted for micro communities, based on window size=1000, for a Diet and COVID-19 dataset [14].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2021 doi:10.20944/preprints202102.0260.v3

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".

Declarations

- Ethics approval and consent to participate: Not Applicable or Yes.
- Consent for publication: Yes.
- Availability of data and materials: Yes, in Data Availability section.
- Competing interests: No.
- Funding: Yes, in Funding section.
- Authors' contributions: The first author was involved in idea generation, text writing, data gathering and processing, chart preparation, and text editing. The second author has been involved in ideation, guidance and supervision, evaluation, text editing and research process management.
- Acknowledgements: Yes, in Acknowledgement section.

Acknowledgement

A preprint has previously been published [15]. Thanks to the comments of Dr. Alireza Talebpour and his fellow researchers.

Data Availability

The IR-QUMA dataset [7] is available from:

https://data.mendeley.com/datasets/d89gphmnsk/3

Funding

Sharif University of Technology, Tehran, Iran, 90300439, student grant for Mohammad Reza Besharati.

References

[1] Agrawal, Prachi, Hattan F. Abutarboush, Talari Ganesh, and Ali Wagdy Mohamed. "Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019)." IEEE Access 9 (2021): 26766-26791.

[2] Jafari, Nafiseh, Mohammad Reza Besharati, Mohammad Izadi, and Maryam Hourali. "SELM: Software Engineering of Machine Learning Models." *arXiv preprint arXiv:2103.11249* (2021).

[3] Chen, Yan, Dan Li, Yanhai Li, Xiaoyuan Ma, and Jianming Wei. "Use moving average filter to reduce noises in wearable PPG during continuous monitoring." In *eHealth 360°*, pp. 193-203. Springer, Cham, 2017.

[4] Gholizadeh, B., Numerical Analysis Methods, Sharif University Press, 2012.

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

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

[7] Besharati, Mohammad Reza; Izadi, Mohammad (2020), "IR-QUMA", Mendeley Data, V3, doi: 10.17632/d89gphmnsk.3

[8] Sigari, Mohamad Hoseyn, Naser Mozayani, and H. Pourreza. "Fuzzy running average and fuzzy background subtraction: concepts and application." International Journal of Computer Science and Network Security 8, no. 2 (2008): 138-143.

[9] Likert, Rensis. "A technique for the measurement of attitudes." *Archives of psychology* (1932).

[10] Zadeh, Lotfi A. "Fuzzy logic." Computer 21, no. 4 (1988): 83-93.

[11] Yang, Haijun. "Measuring software product quality with ISO standards base on fuzzy logic technique." In *Affective Computing and Intelligent Interaction*, pp. 59-67. Springer, Berlin, Heidelberg, 2012.

[12] Castagliola, Philippe. "Approximation of the normal sample median distribution using symmetrical Johnson SU distributions: application to quality control." *Communications in Statistics-Simulation and Computation* 27, no. 2 (1998): 289-301.

[13] Oppenheim, Alan V., and Donald H. Johnson. "Discrete representation of signals." *Proceedings of the IEEE* 60, no. 6 (1972): 681-691.

[14] Mohammad Reza Besharati, Mohammad Izadi, & Alireza Talebpour. (2021). Honey, HoneyBee Venom Melittin and SARS-Cov-2 (Version 2). Zenodo. https://doi.org/10.5281/zenodo.4569108

[15] Besharati, M.R.; Izadi, M. SimulaD: A Novel Feature Selection Heuristics for Discrete Data. Preprints 2021, 2021020260 (doi: 10.20944/preprints202102.0260.v2).