Artificial Data Sets for Missing Value Imputation via Clusterwise Linear Regression

Napsu Karmitsa, Sona Taheri, Adil Bagirov, and Pauliina Mäkinen

***Abstract*—This paper consist description of artificial data sets used in the paper "Missing Value Imputation via Clusterwise Linear Regression".**

***Index Terms*—Artificial Data.**

I. Artificial Data

THE occurrence of missing (or incomplete) data is very common in many fields of research such as social sciences, biology, medicine and climate science. To demonstrate the performance of the proposed IVIACLR in different types of data, four synthetic data sets were generated.

The first data set, D500, has no structure in it. It is generated using the uniform distribution within [-2,2]. The number of data points is 500 and the number of features is 4.

The second data set, U500, has three clearly separated clusters. It contains 500 data points and 2 features. The ranges of the feature values with in each cluster and the numbers of points n_i within the cluster are given in Table I while the data set is illustrated in Figure 1.

TABLE I
U500 DATA.

Feature / Cluster	n_i	1	2
1	100	[0,4]	[1.5,2.5]
2	350	[5,15]	[8,10]
3	50	[15,17]	[3,4]

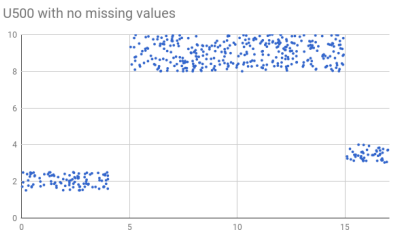

Fig. 1. Original U500

N. Karmitsa and P. Mäkinen are with Department of Mathematics and Statistics, University of Turku, FI-20014 Turku, Finland; e-mail: napsu@karmitsa.fi

S. Taheri and A. Bagirov are with Faculty of Science and Technology, Federation University Australia, Victoria, Australia.

The third and forth data sets U2500 and U10000 consist of 2500 and 10000 data points, respectively, with feature values between 0 and 100. They both have 20 features and 5 clusters. Some of the clusters slightly overlap (i.e. it is not really clear to which clusters some of the points should belong) but the structure of the data is still clear. The more detailed descriptions of these data are given in Tables II and III.

All the artificial data used in our experiments can be found from GitHub.

Acknowledgment

The work was financially supported by the Academy of Finland (Project No. 289500, 294002, and 313266) and Australian Research Counsil's Discovery Projects funding scheme (Project No. DP140103213).

TABLE II
U2500 DATA.

Feature Cluster	n_i	1	2	3	4	5	6	7	8	9	10
1	1000	[0,4]	[5,12]	[20,23]	[30,36]	[38,43]	[44,47]	[49,50]	[52,55]	[57,59]	[60,64]
2	200	[10,11]	[17,19]	[37,40]	[15,17]	[18,20]	[0,4]	[70,72]	[0,1]	[12,13]	[18,20]
3	400	[20,23]	[30,32]	[50,52]	[3,5]	[22,23]	[60,62]	[11,14]	[20,24]	[70,73]	[0,3]
4	400	[50,54]	[50,54]	[90,93]	[40,43]	[70,72]	[50,52]	[20,22]	[70,72]	[30,31]	[85,88]
5	500	[70,72]	[80,82]	[60,61]	[80,81]	[1,2]	[5,6]	[40,42]	[90,92]	[0,2]	[40,43]

Feature Cluster	11	12	13	14	15	16	17	18	19	20
1	[64,65]	[70,76]	[77,80]	[83,88]	[90,96]	[10,15]	[20,25]	[43,48]	[70,72]	[98,100]
2	[22,25]	[37,39]	[1,2]	[5,6]	[60,63]	[21,24]	[17,19]	[80,82]	[80,81]	[90,91]
3	[80,82]	[10,12]	[90,92]	[50,51]	[20,21]	[95,96]	[95,97]	[33,35]	[37,40]	[0,2]
4	[0,3]	[50,52]	[33,35]	[70,71]	[30,31]	[80,81]	[60,61]	[5,7]	[5,7]	[20,24]
5	[30,33]	[90,93]	[15,18]	[20,23]	[70,74]	[50,53]	[0,2]	[10,12]	[90,92]	[50,51]

TABLE III
U10000 DATA.

Feature Cluster	n_i	1	2	3	4	5	6	7	8	9	10
1	4000	[0,4]	[5,12]	[20,23]	[30,36]	[38,43]	[44,47]	[49,50]	[52,55]	[57,59]	[60,64]
2	800	[10,11]	[17,19]	[37,40]	[15,17]	[18,20]	[0,4]	[70,72]	[0,1]	[12,13]	[18,20]
3	1600	[20,23]	[30,32]	[50,52]	[3,5]	[22,23]	[60,62]	[11,14]	[20,24]	[70,73]	[0,3]
4	1600	[50,54]	[50,54]	[90,93]	[40,43]	[70,72]	[50,52]	[20,22]	[70,72]	[30,31]	[85,88]
5	2000	[70,72]	[80,82]	[60,61]	[80,81]	[1,2]	[5,6]	[40,42]	[90,92]	[0,2]	[40,43]

Feature Cluster	11	12	13	14	15	16	17	18	19	20
1	[64,65]	[70,76]	[77,80]	[83,88]	[90,96]	[10,15]	[20,25]	[43,48]	[70,72]	[98,100]
2	[22,25]	[37,39]	[1,2]	[5,6]	[60,63]	[21,24]	[17,19]	[80,82]	[80,81]	[90,91]
3	[80,82]	[10,12]	[90,92]	[50,51]	[20,21]	[95,96]	[95,97]	[33,35]	[37,40]	[0,2]
4	[0,3]	[50,52]	[33,35]	[70,71]	[30,31]	[80,81]	[60,61]	[5,7]	[5,7]	[20,24]
5	[30,33]	[90,93]	[15,18]	[20,23]	[70,74]	[50,53]	[0,2]	[10,12]	[90,92]	[50,51]