Gujarat Journal of Extension Education Vol. 30 : Issue 2 : December 2019

DATA NORMALIZATION IN DATA MINING USING GRAPHICAL USER INTERFACE: A PRE-PROCESSING STAGE

G. J. Kamani[1], R. S. Parmar[2] and Y. R. Ghodasara[3]

1 Asst. Prof., College of Agricultural Information Technology, AAU, Anand -388110
2 Asso. Prof., College of Agricultural Information Technology, AAU, Anand -388110
3 Professor, College of Agricultural Information Technology, AAU, Anand -388110
Email : kamani_gautam@aau.in

ABSTRACT

As we know that the normalization is a technique often applied as component of data preparation for data mining. The objective of normalization is to change the values of numeric attributes in the dataset to a common scale, without altering differences in the ranges of values. Mainly normalization plays vital role for manipulation of data like scale down or scale up the range of data before it becomes used for further analysis. There are different methods of data normalization namely Min-Max normalization, Z-score normalization and Decimal scaling normalization. So by referring these normalization methods we are going to show our proposed user-friendly GUI tool. This GUI tool has the facility to normalize data using Min-Max normalization, Z-score normalization and Decimal scaling normalization methods.

Keywords : *data normialization, data mining*

INTRODUCTION

Data mining techniques are likely to continue as a major driver of agriculture development in India. Data mining in agriculture is an emerging research topic. Data mining is a process of discovering previously unknown patterns used for strategic decision making. There are various steps involved in mining process such as Data Integration, Data Selection, Data Cleaning, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Presentation and Use of Discovered Knowledge. Data normalization is generally required when we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an important equally important attribute(on lower scale) because of other attribute having values on larger scale. In other words, when multiple attributes are there with different scales, this may lead to poor data models during data mining operations. So they are normalized to bring all the attributes on the same scale. Normalization is scaling technique or a mapping technique or a pre processing stage, where, we can find new range from an existing one range (Shalabi *et al.,* 2006). As we know, there are so many ways to predict or forecast but they vary with one another a lot. So to maintain the large variation of prediction and forecasting the normalization technique is required to make them closer. (Sanjaya *et al.,* 2014 and Sanjaya and Prasanta, 2015). Patrol and Sahu (2015) studied Integer Scaling normalization technique, that works well in soft computing ,image processing and cloud computing etc.

METHODOLOGY

GUI (Graphical User Interface) tool to normalize data using Min-Max normalization, Z-score normalization and Decimal scaling normalization methods has been implemented as a layered structure having three layers viz., User Interface Layer (UIL), Application Layer (APL) and Database Layer (DBL). The layer structure of said GUI tool is presented in Fig. 1.

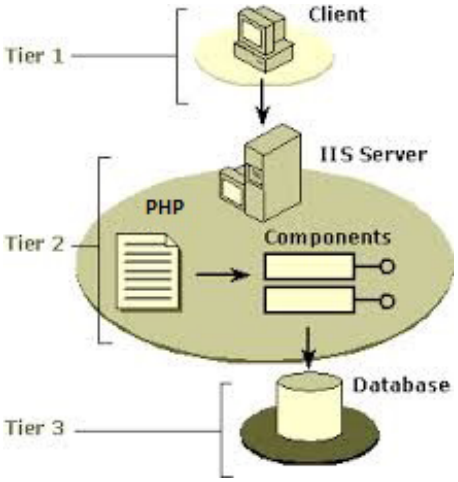

Fig. 1 Layer structure of GUI tool

Methods of data normalization are:

- **Min-Mix normalization**: Transform the data from measured units to a new interval from new_min_A to new_max_A to for feature A .

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

Where v is the respective value of the attribute

v' is Min-Max Normalized data one

min_A is the respective Minimum of value of the attribute

max_A is the respective Maximum value of the attribute

Z-Score normalization

Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, *v*, of A is normalized to *v'* by computing:

$$v' = \frac{v - \bar{F}}{\sigma_F}$$

where and are the mean and standard deviation of feature *F*, respectively.

Decimal scaling

Transform the data by moving the decimal points of values of feature *F*. The number of decimal points moved depends on the maximum absolute value of *F*. A value *v* of *F* is normalized to *v'* by computing :

$$v' = \frac{v}{10^j},$$

Where *j* is the smallest integer such that *Max v'* <1

RESULTS AND DISCUSSION

The Home Page (Fig.2) of GUI tool has "**Upload File**", button. By clicking this button one can upload the selected XLS file.

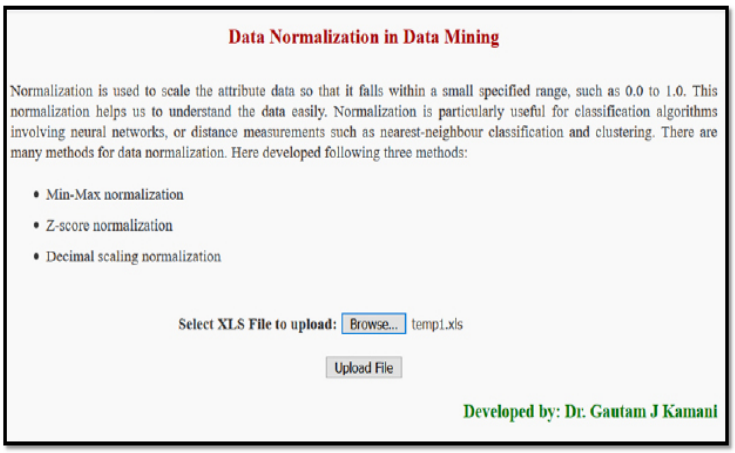

Fig. 2 Home Page

The Normalization Method Page (Fig.3) has options like "Field Label", "Normalization Method", and "Range for Min-Max Method". By clicking Normalization Method list box one can select the desired normalize method namely Min-Max, Decimal Scale or Z-Score for respective Field Label. If one can select the Min-Max method, then it is required to enter the range for Min-Max Method. By clicking on Process Button one can get the desired normalize data.

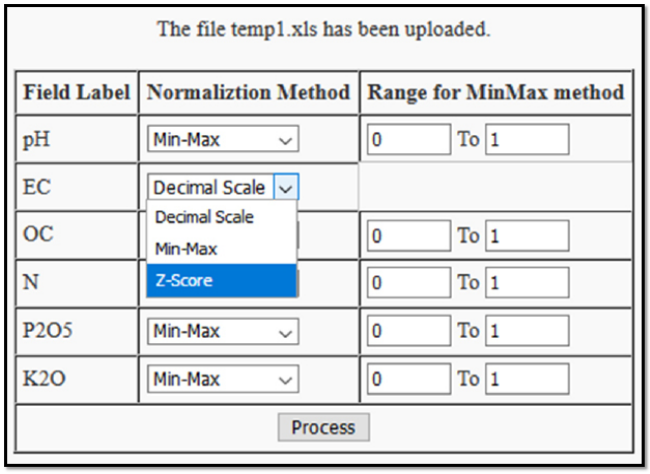

Fig. 3 Normalization Method Page

Fig.4, Fig.5 and Fig.6 show the results of normalized data using Min-Max normalization, Z-score normalization and Decimal Scaling normalization methods respectively.

Field Lable	Min Value	Max Value	Abs. Max Value	Standard Deviation
pH	6.75	8.44	8.44	0.44386246618619
EC	0.19	0.54	0.54	0.11926342872859
OC	0.15	0.97	0.97	0.1629587541544
N	125	282	282	41.964013095291
P2O5	30.513277822608	376.52418158056	376.52418158056	78.613623161891
K2O	171.024	1091.664	1091.664	193.70611059547

pH		EC		OC		N		P2O5		K2O	
Original	Min-Max (0-1)	Original	Min-Max (0-1)	Original	Min-Max (0-1)	Original	Min-Max (0-1)	Original	Min-Max (0-1)	Original	Min-Max (0-1)
6.75	**0**	0.19	**0**	0.62	**0.5732**	172	**0.2994**	35.54	**0.0145**	301.532	**0.1418**
7.4	**0.3846**	0.34	**0.4286**	0.59	**0.5366**	251	**0.8025**	66.446718765295	**0.1039**	430.416	**0.2818**
7.65	**0.5325**	0.3	**0.3143**	0.51	**0.439**	125	**0**	30.513277822608	**0**	446.544	**0.2993**
8.07	**0.7811**	0.49	**0.8571**	0.68	**0.6463**	282	**1**	163.76132150629	**0.3851**	399.504	**0.2482**
8.19	**0.8521**	0.54	**1**	0.54	**0.4756**	188	**0.4013**	88.100194818195	**0.1664**	430.416	**0.2818**
8.44	**1**	0.3	**0.3143**	0.29	**0.1707**	157	**0.2038**	98.711153389991	**0.1971**	321.552	**0.1635**
7.68	**0.5503**	0.32	**0.3714**	0.39	**0.2927**	188	**0.4013**	156.97638567265	**0.3655**	382.032	**0.2292**
8.35	**0.9467**	0.23	**0.1143**	0.53	**0.4634**	157	**0.2038**	179.03737970056	**0.4292**	383.376	**0.2307**
7.97	**0.7219**	0.41	**0.6286**	0.55	**0.4878**	125	**0**	169.25973109524	**0.401**	597.072	**0.4628**
7.9	**0.6805**	0.52	**0.9429**	0.6	**0.5488**	125	**0**	132.04067965209	**0.2934**	494.928	**0.3518**
7.91	**0.6864**	0.38	**0.5429**	0.53	**0.4634**	125	**0**	234.34362155941	**0.5891**	661.584	**0.5328**
8.31	**0.9231**	0.24	**0.1429**	0.4	**0.3049**	157	**0.2038**	126.95337423963	**0.2787**	320.208	**0.162**
7.84	**0.645**	0.53	**0.9714**	0.59	**0.5366**	157	**0.2038**	116.93893034474	**0.2498**	474.768	**0.3299**
8.01	**0.7456**	0.52	**0.9429**	0.53	**0.4634**	188	**0.4013**	128.85475177914	**0.2842**	348.432	**0.1927**
8.1	**0.7988**	0.26	**0.2**	0.53	**0.4634**	125	**0**	61.459296847196	**0.0894**	171.024	**0**
7.9	**0.6805**	0.27	**0.2286**	0.48	**0.4024**	157	**0.2038**	98.11541292318	**0.1954**	365.904	**0.2117**
7.99	**0.7337**	0.49	**0.8571**	0.97	**1**	188	**0.4013**	376.52418158056	**1**	1091.664	**1**
6.87	**0.071**	0.22	**0.0857**	0.15	0	188	**0.4013**	82.305787087335	**0.1497**	265.104	**0.1022**

Fig.4 Normalize Data using Min-Max Normalization Method

Field Lable	Min Value	Max Value	Abs. Max Value	Standard Deviation
pH	6.75	8.44	8.44	0.44386246618619
EC	0.19	0.54	0.54	0.11926342872859
OC	0.15	0.97	0.97	0.1629587541544
N	125	282	282	41.964013095291
P2O5	30.513277822608	376.52418158056	376.52418158056	78.613623161891
K2O	171.024	1091.664	1091.664	193.70611059547

pH		EC		OC		N		P2O5		K2O	
Original	Zscore	Original	Zscore	Original	Zscore	Original	Zscore	Original	Zscore	Original	Zscore
6.75	**-2.482**	0.19	**-1.458**	0.62	**0.5727**	172	**0.0543**	35.54	**-1.2057**	301.532	**-0.7051**
7.4	**-1.0176**	0.34	**-0.2003**	0.59	**0.3886**	251	**1.9368**	66.446718765295	**-0.8126**	430.416	**-0.0397**
7.65	**-0.4543**	0.3	**-0.5357**	0.51	**-0.1023**	125	**-1.0657**	30.513277822608	**-1.2697**	446.544	**0.0435**
8.07	**0.4919**	0.49	**1.0574**	0.68	**0.9409**	282	**2.6756**	163.76132150629	**0.4253**	399.504	**-0.1993**
8.19	**0.7622**	0.54	**1.4767**	0.54	**0.0818**	188	**0.4356**	88.100194818195	**-0.5371**	430.416	**-0.0397**
8.44	**1.3255**	0.3	**-0.5357**	0.29	**-1.4523**	157	**-0.3032**	98.711153389991	**-0.4022**	321.552	**-0.6017**
7.68	**-0.3868**	0.32	**-0.368**	0.39	**-0.8387**	188	**0.4356**	156.97638567265	**0.339**	382.032	**-0.2895**
8.35	**1.1227**	0.23	**-1.1226**	0.53	**0.0205**	157	**-0.3032**	179.03737970056	**0.6196**	383.376	**-0.2826**
7.97	**0.2666**	0.41	**0.3866**	0.55	**0.1432**	125	**-1.0657**	169.25973109524	**0.4952**	597.072	**0.8206**
7.9	**0.1089**	0.52	**1.309**	0.6	**0.45**	125	**-1.0657**	132.04067965209	**0.0218**	494.928	**0.2933**
7.91	**0.1314**	0.38	**0.1351**	0.53	**0.0205**	125	**-1.0657**	234.34362155941	**1.3231**	661.584	**1.1537**
8.31	**1.0326**	0.24	**-1.0388**	0.4	**-0.7773**	157	**-0.3032**	126.95337423963	**-0.0429**	320.208	**-0.6087**
7.84	**-0.0263**	0.53	**1.3928**	0.59	**0.3886**	157	**-0.3032**	116.93893034474	**-0.1703**	474.768	**0.1892**
8.01	**0.3567**	0.52	**1.309**	0.53	**0.0205**	188	**0.4356**	128.85475177914	**-0.0187**	348.432	**-0.463**
8.1	**0.5595**	0.26	**-0.8711**	0.53	**0.0205**	125	**-1.0657**	61.459296847196	**-0.876**	171.024	**-1.3788**
7.9	**0.1089**	0.27	**-0.7872**	0.48	**-0.2864**	157	**-0.3032**	98.11541292318	**-0.4097**	365.904	**-0.3728**
7.99	**0.3117**	0.49	**1.0574**	0.97	**2.7205**	188	**0.4356**	376.52418158056	**3.1317**	1091.664	**3.3739**
6.87	**-2.2116**	0.22	**-1.2065**	0.15	**-2.3114**	188	**0.4356**	82.305787087335	**-0.6108**	265.104	**-0.8932**

Fig.5 Normalize Data using Z-score Normalization Method

Field Lable	Min Value	Max Value	Abs. Max Value	Standard Deviation
pH	6.75	8.44	8.44	0.44386246618619
EC	0.19	0.54	0.54	0.11926342872859
OC	0.15	0.97	0.97	0.1629587541544
N	125	282	282	41.964013095291
P2O5	30.513277822608	376.52418158056	376.52418158056	78.613623161891
K2O	171.024	1091.664	1091.664	193.70611059547

pH		EC		OC		N		P2O5		K2O	
Original	Decimal	Original	Decimal	Original	Decimal	Original	Decimal	Original	Decimal	Original	Decimal
6.75	**0.000675**	0.19	**1.9E-5**	0.62	**6.2E-5**	172	**0.172**	35.54	**3.554E-14**	301.532	**3.01532E-6**
7.4	**0.00074**	0.34	**3.4E-5**	0.59	**5.9E-5**	251	**0.251**	66.446718765295	**6.6446718765295E-14**	430.416	**4.30416E-6**
7.65	**0.000765**	0.3	**3.0E-5**	0.51	**5.1E-5**	125	**0.125**	30.513277822608	**3.0513277822608E-14**	446.544	**4.46544E-6**
8.07	**0.000807**	0.49	**4.9E-5**	0.68	**6.8E-5**	282	**0.282**	163.76132150629	**1.6376132150629E-13**	399.504	**3.99504E-6**
8.19	**0.000819**	0.54	**5.4E-5**	0.54	**5.4E-5**	188	**0.188**	88.100194818195	**8.8100194818195E-14**	430.416	**4.30416E-6**
8.44	**0.000844**	0.3	**3.0E-5**	0.29	**2.9E-5**	157	**0.157**	98.711153389991	**9.8711153389991E-14**	321.552	**3.21552E-6**
7.68	**0.000768**	0.32	**3.2E-5**	0.39	**3.9E-5**	188	**0.188**	156.97638567265	**1.5697638567265E-13**	382.032	**3.82032E-6**
8.35	**0.000835**	0.23	**2.3E-5**	0.53	**5.3E-5**	157	**0.157**	179.03737970056	**1.7903737970056E-13**	383.376	**3.83376E-6**
7.97	**0.000797**	0.41	**4.1E-5**	0.55	**5.5E-5**	125	**0.125**	169.25973109524	**1.6925973109524E-13**	597.072	**5.97072E-6**
7.9	**0.00079**	0.52	**5.2E-5**	0.6	**6.0E-5**	125	**0.125**	132.04067965209	**1.3204067965209E-13**	494.928	**4.94928E-6**
7.91	**0.000791**	0.38	**3.8E-5**	0.53	**5.3E-5**	125	**0.125**	234.34362155941	**2.3434362155941E-13**	661.584	**6.61584E-6**
8.31	**0.000831**	0.24	**2.4E-5**	0.4	**4.0E-5**	157	**0.157**	126.95337423963	**1.2695337423963E-13**	320.208	**3.20208E-6**
7.84	**0.000784**	0.53	**5.3E-5**	0.59	**5.9E-5**	157	**0.157**	116.93893034474	**1.1693893034474E-13**	474.768	**4.74768E-6**
8.01	**0.000801**	0.52	**5.2E-5**	0.53	**5.3E-5**	188	**0.188**	128.85475177914	**1.2885475177914E-13**	348.432	**3.48432E-6**
8.1	**0.00081**	0.26	**2.6E-5**	0.53	**5.3E-5**	125	**0.125**	61.459296847196	**6.1459296847196E-14**	171.024	**1.71024E-6**
7.9	**0.00079**	0.27	**2.7E-5**	0.48	**4.8E-5**	157	**0.157**	98.11541292318	**9.811541292318E-14**	365.904	**3.65904E-6**
7.99	**0.000799**	0.49	**4.9E-5**	0.97	**9.7E-5**	188	**0.188**	376.52418158056	**3.7652418158056E-13**	1091.664	**1.091664E-5**
6.87	**0.000687**	0.22	**2.2E-5**	0.15	**1.5E-5**	188	**0.188**	82.305787087335	**8.2305787087335E-14**	265.104	**2.65104E-6**

Fig.6 Normalize Data using Decimal Scaling Normalization Method

CONCLUSION

The use of computers has changed the whole complexion of research. Computers play a very important role in Agricultural statistics. With the development of Statistical software packages, the statistical analysis of data has become relatively easy. Now it is possible to carry out sophisticated data analysis in no time. Experimentation and making conclusion are twin essential features of general scientific methodology. Statistics and Computer Science discipline are mainly design to achieve these objectives. Data Mining Techniques are often used by the researcher. Data normalization is required when we are dealing with attributes on a different scale. We use different methods of data normalization namely Min-Max normalization, Z-score normalization and Decimal scaling normalization. This GUI tool has the facility to normalize data using Min-Max normalization, Z-score normalization and Decimal scaling normalization methods. This tool is made very simple to use.

REFERENCES

Patro1, S. G. and Sahu, K. K. (2015) Normalization: A Pre-processing Stage, *International Advanced Research Journal in Science, Engineering and Technology* (IARJSET).

Raj, M. P., Vegad, N. M. and Amin, B. A. (2018) Big data: smart agriculture. *Guj. J. Ext. Edu.* 29(2):245-247.

Sanjaya, K. P.; Subhrajit, N. and Prasanta, K. J. (2014). A Smoothing Based Task Scheduling Algorithm for Heterogeneous Multi-Cloud Environment, 3rd IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, Waknaghat, 11th - 13th Dec 2014.

Sanjaya, K. P. and Prasanta, K. J. (2015)., Efficient Task Scheduling Algorithms for Heterogeneous Multi-cloud *Environment, The Journal of Supercomputing,* Springer.

Shalabi, L.A.; Shaaban, Z. and Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine, *Journal of Computer Sci.,* 2: 735-739

Vinaya Kumar, H. M., Shivamurthy, M. and Govinda Gowda, V. (2015). Working out norms of distribution of climate-induced crisis management scores by using constructed scale. *Annals of Plant and Soil Research.* 17: 384-386.

Received : September 2019 : Accepted : November 2019