Data Transformation Techniques for Academic Datasets

V.Sathya Durga, Thangakumar Jeyaprakash

Abstract*: Data mining is a real-world procedure of discovering useful patterns from heterogeneous datasets. All most all industry uses data mining in their day to day activities. To build an effective mining model, a series of development steps are to be followed. It starts with discovering the business problem and ends with communicating the results. In this development life cycle, the most important step is data preparation or data preprocessing. Data preprocessing is converting raw data into data understandable by the machine. Data normalization is a phase in data preprocessing where the data values are scaled to 0 and 1. Right normalization of the datasets leads to improved mining results. In this paper, academic data of students is taken. The dataset is normalization using six normalization technique. Multi Layer Perceptron classifier is applied to normalized dataset and results are obtained. Results of this study reveal the best normalization technique which can be used for normalizing academic datasets. Finally, in a line, the goal of this work is to discover the best normalization technique which produces better mining result when applied to academic datasets.

***Keywords*: Cube Root Normalization; Data Normalization; Decimal Scaling Normalization; Root Mean Square Error**

I. INTRODUCTION

Data mining is new science of finding out novel trends from data. From the discovered patterns, the set of rules are derived. These derived rules enable users to take timely decisions and to make an in-depth study of the data. Vast amount of data is generated by all business units and by major industries like health care, retail, manufacturing, etc. Health industry use data mining to predict whether a patient will be diabetics or not in the next ten years, the retail industry use data mining to analyze the shopping patterns of the customers [1]. Manufacturing industry uses data mining for industrial process automation.

Applications of data mining may vary from industry to industry but the essence of all these applications remains the same, which is to provide better decision making capabilities to the end users. Development of these high end data mining systems with end to end decision capabilities are built by a typical data mining life cycle. It is a six phase development process which starts with the discovery of the business problem and ends with communicating the results to the end users. The next section discusses data mining life cycle model briefly which is used to develop mining applications.

II. UNDERSTANDING DATA MINING

A. Typical Data Mining Life Cycle

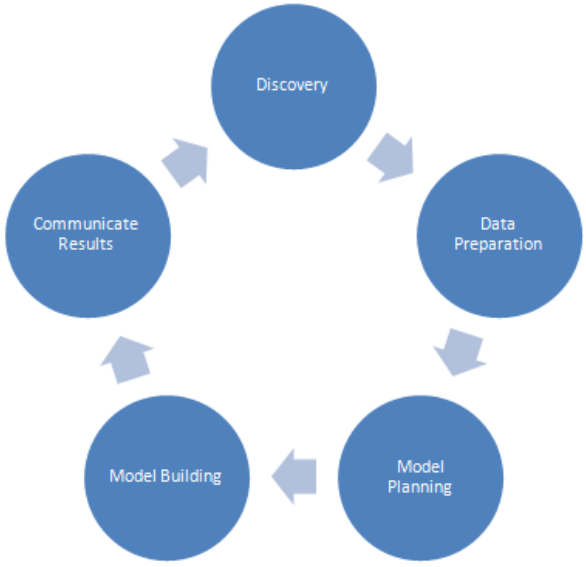

Fig. 1. (a) Typical Data Mining Lifecycle

Discovery: In the discovery phase business problem is identified, initial hypothesis to be tested is framed and nature of the data and the dataset is learned

Data Preparation: In this phase that data to be modeled is explored, preprocessed and conditioned prior to planning and building phase.

Model Planning: In this model planning phase, methods used to develop the data mining model, the techniques used for development and the underlying workflow is determined.

Model building: This is a work intense phase where separate datasets are developed for testing and training. Separate dataset is produced for the production process. In the model building phase the data mining model is built and results are obtained.

Communicate results: This is the final phase where the key finding is identified, the business value is quantified, narratives are summarized and key findings are conveyed to the stake holders.

Revised Manuscript Received on October 15, 2019.

V.Sathya Durga, Research Scholar, Department of CSE, Hindustan Institute of Technology and Science, Padur, India.

Thangakumar Jeyaprakash, Associate Professor, Department of CSE, Hindustan Institute of Technology and Science, Padur, India.

B. Data Preprocessing and Data Normalization

The most significant and the critical step in the development lifecycle of any data mining application is data preprocessing. It is a known fact that real-world data is incomplete and inconsistent. Before we train and build the model with data, the data has to be preprocessed for better performance and to achieve reliable results. Data preprocessing is a detailed and a stepwise process which begins with data cleaning, integration, transformation and which ends with data reduction[3].

Out of the above mentioned data preprocessing activities, data normalization is the most important part of data preprocessing . In data normalization, data values are scaled to the range of 0.0 to 0.1 normally. Normalization helps to reduce the misclassification error in classification problems and to reduce the root mean square error in regression problems [4].

In this work, we have taken academic datasets and the values in the data set is normalization using six normalization techniques. Machine Learning Classifier the Multi Layer Perceptron is applied and the results are tabulated. This paper intends to identify the most efficient transformation techniques for academic datasets. The next section discusses the works of other researchers on normalization.

C. Related Work

Gopal Krishna and Kishore Kumar propose a technique namely Integer Scaling normalization. The proposed technique was tested on three datasets. It is noted that the proposed technique works effectively and produces a range of values between 0 and 1. It is quoted that this normalization technique works well in soft computing and cloud computing [5].

Saranya and Manikandan experiment and compares three normalization techniques in their work. The task taken is to use normalization for privacy preservation. Age attribute in the dataset is normalized. The normalized dataset is tested with K Means clustering. Results reveal that Min Max Normalization outperforms other two techniques [6].

Zuriani and Yuhanis investigate three normalization techniques to predict the outbreak of dengue fever. Three datasets, two relating to dengue and one rainfall dataset is used in this work. Data were normalized using the three normalization techniques. To the normalized datasets, machine learning classifiers were applied. Experimental values prove that decimal point value achieves maximum accuracy of 86.84% [7].

III. MATERIALS AND METHODS

This section explains the data set used in this work, the attributes in the data sets, the transformation technique used, the classifiers applied and the metrics used to evaluate the results of the work.

A. Dataset Used

The dataset used in this paper contains 395 records of math's mark of two schools from Portuguese [8]. The dataset used in this research work is a standard dataset downloaded from the UCI repository. This standard dataset has been used by many researchers in their research works.

B. Attributes in the Datasets

The dataset contains 33 attributes. The dataset constitutes students personal details, academic details, socio economic details. The most important attributes of this dataset are the grades of students in three exams. Some of the interesting attributes identified from this datasets are family size, parents job, study time, travel time, internet at home, etc.

C. Transformation Techniques

The transformation techniques used in this research work are as follows [9]

- Cube root Normalization
- Decimal Scaling Normalization
- Min - Max Normalization
- Normalization using Norm
- Square root Normalization
- Z Score Normalization

D. Classifiers Applied

The learning classifier applied is Multi Layer Percepton.

E. Metrics used for evaluation

The metrics used for evaluation in this research work is Root Mean Square Error (RMSE). RMSE denotes the value difference between the actual and the predicted value.

F. Methodology followed

The methodology used in this study is, first the dataset downloaded from the UCI repository is examined. Then data is normalized with one of the six techniques mentioned above. The dataset is split with various ratios [10]. Next, the ML classifier is applied to the normalized datasets. Results obtained are compared to the standard evaluation metrics.

G. Software Used

Matlab 2018 a, R Studio and Weka 3.8 was the software used in this research work

IV. EXPIREMENTAL RESULTS

The results of the normalization experiments are tabulated and given below. Table 1 lists the RMSE of the academic data set in various split ratios after applying the MLP Classifier. Table 2 tabulates time required to construct the model with normalized data with various split conditions. Figure 1,2 and 3 compares the Root Mean Square Error values of the normalization technique with each split condition in a bar plot.

Table 1.RMSE Values

Split	*Cube root transformation*	*Decimal Scaling*	*Min Max*	*Norm Values*	*Square root*	*Z Score*
90:10	0.9843	0.0963	5.83	0.9507	4.816	9.632
80:20	0.9196	0.101	5.87	0.8775	5.0475	10.095
70:30	0.7576	0.1181	5.96	0.8489	5.9026	11.8051
60:40	0.8214	0.1187	5.96	1.0171	5.9353	11.8706
50:50	0.9163	0.1305	5.86	0.829	6.5787	13.0681
40:60	0.9499	0.1282	5.82	1.0493	6.4083	12.8165
30:70	1.0859	0.132	5.87	0.8939	6.6015	13.203
20:80	1.0483	1.1321	5.87	0.8452	6.6051	13.2102
10:90	1.1295	0.1468	5.84	1.1242	7.3378	14.6756

Table 2. Time taken to build the model

Split	*Cube root transformation*	*Decimal Scaling*	*Min Max*	*Norm Values*	*Square root*	*Z Score*
90:10	5.83	6.12	5.75	5.87	5.98	5.72
80:20	5.87	5.71	5.73	5.84	5.93	5.65
70:30	5.96	5.74	5.77	5.83	5.92	5.85
60:40	5.96	5.84	5.83	5.85	5.96	5.65
50:50	5.86	5.88	5.76	5.83	5.92	5.76
40:60	5.82	5.81	5.72	5.86	5.9	5.62
30:70	5.87	5.99	5.71	6.06	5.9	5.63
20:80	5.87	5.84	5.75	5.91	5.91	5.65
10:90	5.84	5.99	5.74	5.92	5.86	5.6

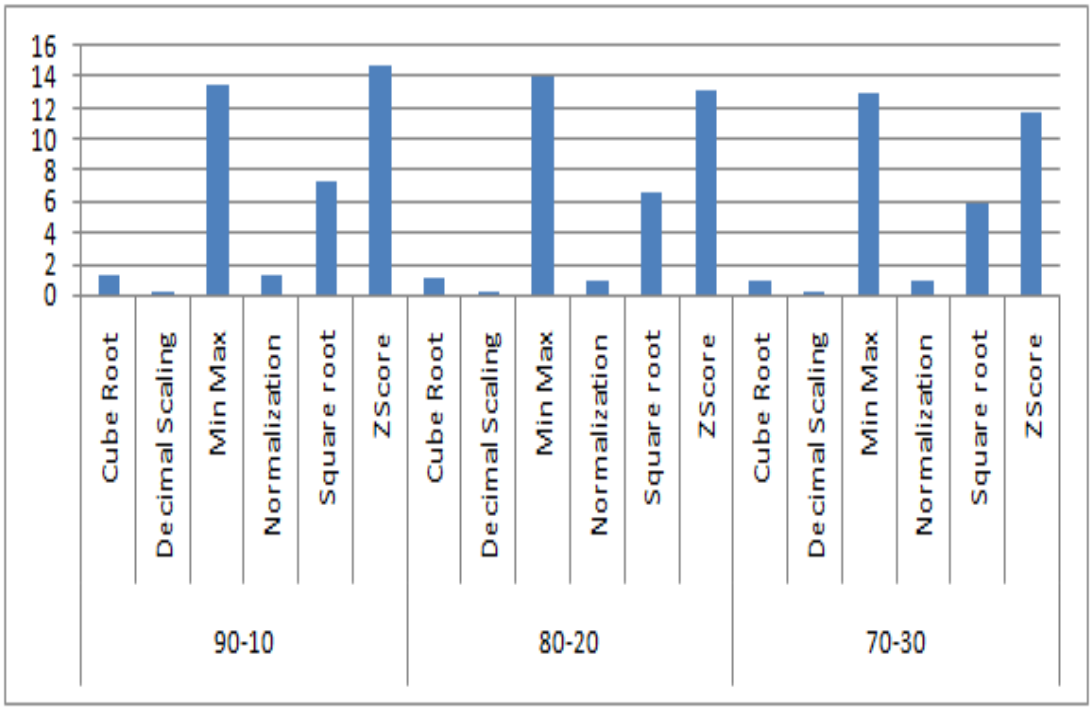

Fig. 2. RMSE Values of 90-10, 80-20 and 70-30 Split.

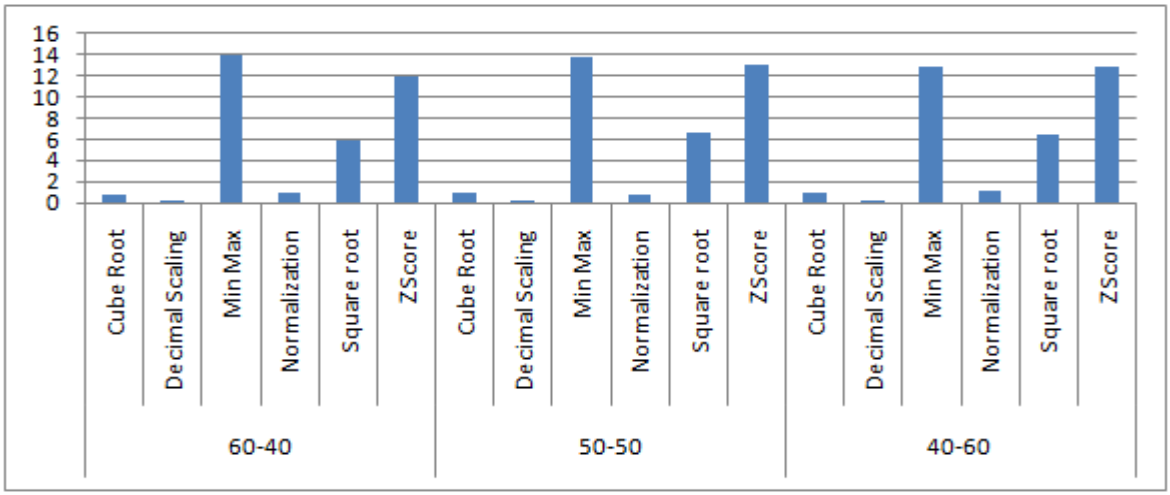

Fig. 3. RMSE Values of 60-40, 50-50 and 40-60 Split

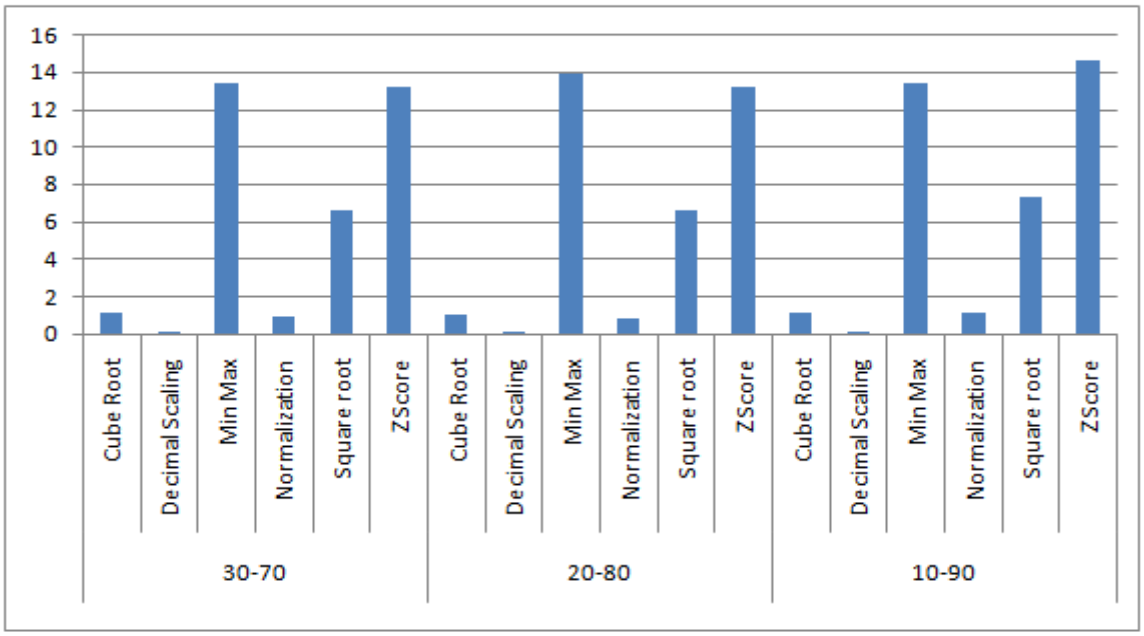

Fig. 4. RMSE Values of 30-70, 20-80 and 10-90 Split.

V. EXPIREMENTAL RESULTS

A. Findings 1

- Results of cube root transformation revel that Root mean square error is less in 70:30 split.
- Results of decimal scaling transformation shows that Root mean square error is less in 80:20 split.
- Results of Min-Max transformation produces a low Root mean square error at 40:60 split.
- Results of Normalization by norm value shows that Root mean square error is less in 50:50 split.
- Results of square root Normalization revels that Root mean square error is less in 90:10 split.
- Results of square root Normalization shows that Root mean square error is less in 90:10 split.

B. Findings 2

- Experimental results show that overall low Root Mean Square Error is achieved by Decimal scaling normalization with the lowest value being 0.101, at 80:20 split.
- Cube root transformation achieves an RMSE of 0.7576 at 70:30 split, making it the second best choice for normalization.
- The next best result is produced with Normalization by norm values with 0.829 RMSE at 50:50 split.

C. Findings 3

- Z score normalization achieves lowest time to build the model with 5.6 seconds.
- Decimal scaling normalization takes the highest time to build the model with 6.12 seconds.

VI. DISCUSSION AND CONCLUSION

A. Discussions

In some literature, it was reported that Min Max normalization performs better than other normalization techniques in transforming data. But the results of this research works show that for academic datasets Min Max technique fails to achieve low root mean square error. From the experimental results of this work, we can comprehend that for academic data set the best technique to transform data is by decimal scaling, followed by cube root normalization.
It is inferred, that time taken to build the model is high in decimal scaling compared to other techniques. But as academic systems are not rigid time bound systems giving high priority to latency and time delays, we can look into the low RMSE of decimal scaling normalization which is more important for any prediction model rather than the time taken to build the model.

B. Conclusion

In this paper, we have taken academic datasets, normalized these datasets using six normalization technique. Using the MLP classifiers value of RMSE is obtained. Obtained results are tabulated and presented. From the results, we can conclude that for academic dataset the best data transformation technique is decimal scaling, followed by cube root normalization.

REFERENCES

1. Role of Data Mining in retail sector, .coolavenues.com, http://www.coolavenues.com/marketing-zone/role-of-data-mining-in-retail-sector (accessed May 28, 2019).
2. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. 2015. EMC Education Services, 1st Edition, Chapter 1, Pg. 25.
3. "Data Preprocessing Techniques for Data Mining", http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf, performance (accessed May 28, 2019).
4. Data Warehousing and Data Mining. 2012. ITL Educational Solutions Ltd. First Edition. Chapter 6.pp.111.
5. Gopal Krishna Patro and Kishore Kumar Sahu (2015) "Normalization: A Preprocessing Stage" IARJSET. arXiv:1503.06462.
6. C. Saranya and G.Manikandan (2013) "A study on normalization techniques for privacy preserving data mining", *Int.J.Eng and Tech*, 5(3): 2701-2704.
7. Z. Mustaffa and Y. Yusof. (2011) "A comparison of normalisation techniques in predicting dengue outbreak", Proc. International conference on business and economics research, IACSIT Press: 345-349.
8. "Student Performance Data Set", archive.ics.uci.edu https://archive.ics.uci.edu/ml/datasets/student+performance, (accessed May 28, 2019).
9. "Data transformation—Skewness, normalization and much more. medium.com". https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55, (accessed May 28, 2019).
10. Shabib Aftab, Munir Ahmed, Nourneen Hameed, Muhammad Salam Bashir, Iftikhar Ali, Zahid Nawaz. (2018) "Rainfall Prediction in Lahore City using using Data Minig ". *International Journal of Advanced computer science and Application* **9**(4) :254-260.