



Normalization of Data in Data Mining

Dr. Himani Goyal¹, Sandeep.D²Venu.R³ Raghavendra Pokuri⁴, Sandeep Kathula⁵, Naveen Battula⁶

¹Dean, Dept. of Electronics and Communications, ^{2,3,4,5,6} Student
MLR Institute of Technology, Dundigal, Hyderabad-43 Telangana, India

Abstract: In today's competitive world that thrives on the thirst for profits through excellence, obtaining greater efficiency by optimum utilization of resources and better decision-making through analytical data mining methods has become the backbone of every industry. This highlights the importance of the powerful tools and concepts of Data Mining and Warehousing, which when applied effectively can revolutionize the face of any industry. The role of *normalization* techniques has become extremely pivotal for identifying patterns and maintaining the consistency of database.

Keywords: Data Normalization, Min-Max, Decimal Scaling, Zero-Score.

I. Introduction

Normalization is a process of decomposing the attribute values so that they are within a specified range of smaller size. It transforms a complex database into a simple database. Normalization involves a sequence of rules to be employed to test individual relations so that the database can be normalized to any degree. The process of normalization is based on the engrossing concept of normal forms. A relational schema may be in either 1NF or 2NF or 3NF or Boyce-Codd Normal form. If the relational schema is not in the required normal form, then it has to be transformed into either of the desired normal forms. Normalization can thus be used as a *data transformation technique*. The various data normalization techniques are as follows:

II. Min-Max Normalization

This intriguing technique is responsible for accomplishing linear transformation on actual data set and for retaining the correlation between them. Assume 'R' to be an attribute of a given relational schema. Also, assume that the range of values which 'R' can take may vary from M_p to X_p . In this enticing technique, a value 'd' of attribute R is mapped to d' in the range $[nX_p, nM_p]$ by calculating d' using the equation:

$$d' = (d - M_p)(nX_p - nM_p) / (X_p - M_p) + nM_p$$

An error "out-of-bound" is displayed in computer executed program if the input value is greater than the actual data range.

III. Zero-Score (Z-Score) Normalization:

This method is generally used when the actual range of a particular attribute is unknown. However, this technique can be used to obtain feasible results if the minimum and maximum values are considered to be outliers. Normalization can thus be performed by using arithmetic mean and standard deviation. Thus, the value d may be transformed in d' using the equation:

$$d' = (d - P_A) / \sigma_P$$

Where P_A is the arithmetic mean of attribute P, whereas σ_P is the attribute P.

IV. Normalization using Decimal Scaling

The data value of attribute P is normalized by changing the position of decimal points. The decision regarding the position of decimal point is based on maximum absolute value of P i.e., $\text{Max}(|d'|)$. The value of d is thus transformed using the equation, $d' = d / 10^Z$



V. Elimination of Outliers

Outliers are a common sighting while dealing with data. Their presence creates quite a lot of hassles in the computations. So, eliminating them is a very clever idea. So, detect the outliers from the box-plots and refine the data by eliminating them. One legitimate reason to remove outliers is to prevent the distortion of central

tendency of data. Suppose that the data for analysis includes the attribute age. The age values for the data tuples in the increasing order are 13,15,16,16,19,20,21,22,25,25,25,25,30,33,33,35,35,35,36,40,45,46,52,70.

Thus using the concept of min-max normalization to transform the value 35 for age within the range [0.0,1.0]:

(M_p) min=13 and (X_p) max=70. Range is [nM_p, nX_p]=[0.0,1.0].

Transforming the value 35 as,

$$d' = (d - M_p)(nX_p - nM_p) / (X_p - M_p) + nM_p$$

$$= (35 - 13)(1.0 - 0.0) / (70 - 13) + 0.0$$

$$= 22(1.0) / 57 = 0.38.$$

Hence, $d' = 0.38$ which is well within the actual range. The arithmetic mean $P_A = 29.96$ and Standard deviation $\sigma_p = 12.94$ years. Thus using z-score normalization, $d' = d - P' / \sigma_p$ which is same as $(5.04) / (12.94) = 0.38$. The value obtained using min-max normalization is same as the score obtained through z-score normalization. Further, the value d' can be transformed using decimal scale normalization as, $d' = d / 10^Z = 35 / 10^2 = 0.35$. The value d' is thus approximately 0.365 which is obtained by taking into consideration the mean of the above three values.

V. Application

Normalization is extensively used in the following applications:

- (i) Neural network classification algorithms such as in back-propagation algorithm that enhances the speed of learning phase.
- (ii) Distance-based method such as k-nearest neighbor classification that prohibits the larger range attribute values from outweighing the smaller range attribute values.

VI. Conclusion

Normalized relation tables do not contain repeated groups. Hence the concepts of anomalous updation, anomalous deletion, anomalous insertion, redundancy errors and database inconsistency can be obviated. Further, simplified results can be obtained which help in efficient maintenance of database integrity. Business enterprises can thus enhance their data analytics through the predictive behavior of the normalized data.

Acknowledgments

Ineffable are our feelings to Prof. Kamakshi Prasad of JNTU-Hyderabad for assisting us in this work. The values and beliefs that our professors have instilled in us have been a source of constant inspiration. F.A likes to extend special thanks to his parents for their amazing insight and guidance. The unflinching support of family members through thick and thin has helped us in reaching where we are today. S.A would like to thank the students of JNTU for their constant motivation. T.A extends his warm regards to his friends for their excellent ideologies and ideations which have been the constant sources of enlightenment.

References

1. Database Management Systems by Raghu RamaKrishna, 2002 edition, McGraw -Hill.
2. Database Management Systems by Abraham Silberschatz, Henry F.Korth, Sudarshan, Ed 5, 2005, Mc-Graw-Hill education.
3. Database Management Systems by Raghuram and RadhaKrishna, Professional Publications.
4. Data Mining: Concepts and Techniques by Jiawei Han and Micheline Kamber, Ed4, Morgan Kauffman publications.
5. Data Mining tutorial, tutorialspoint.com.
6. Data Mining Techniques by Arun K Pujari.
7. Fundamentals of Database Systems by Ramez Elmasri & Shamkant Navathe, Ed 4.

About the Authors

First Author: Raghavendra Pokuri is a final year Computer Science Engineering student pursuing his Bachelors in Technology from JNTU-Hyderabad. His fields of interests include research on the captivating subjects of Data Mining and Data Warehousing, adhoc -sensor networks and extensive C& Java programming.

Second Author: Sandeep Kathula is a final year Computer Science engineering student pursuing his Bachelors in Technology from the esteemed college of JNTU-Hyderabad. His fields of interest include extensive research on Information retrieval systems, Data Mining and Warehousing, SQL programming and Web programming.

Third Author: Naveen Battula is a final year Computer Science Engineering student pursuing his Bachelors in Technology from JNTU-Hyderabad. His fields of interest include substantial research on Database Transactions and Concurrency control, Storage and Indexing algorithms, Schema Refinement and Relational Calculus.