Survey on Data Preprocessing Concept Applicable in Data Mining

Mathew Ngwae Maingi[1]

[1]Jomo Kenyatta University of Agriculture and Technology, School of Computing and Information Technology,
P.O. Box 62000-00200 Nairobi, Kenya

Abstract: ***Real world data is highly prone to outliers commonly known as data noise. This occurrence usually causes a problem of missing values or maybe data full of inconsistencies thus resulting to a poor quality data. Poor quality data is unreliable and fake since it never upholds data integrity issues. Principally, computer users wish to harvest data that is reliable and of high integrity and that's where the concept of data preprocessing comes in since quality decisions are directly proportional to quality data. Data preprocessing deals with data preparation and data transformation, and seeks to improve the overall process of data mining and at the same time make the process of knowledge discovery more efficient. This paper therefore focuses on surveying different data preprocessing techniques as used in data mining, exhaustively outlining their major purposes in knowledge discovery process.***

Keywords: Data, noise, integrity, preprocessing, transformation**.**

1. Introduction

Real world data usually contain outliers, sometimes incomplete or inconsistent in nature. Data gathering methods quit often result in problematic occurrences like: giving out of range values that are very difficult to work with, data values that contain unworkable combinations, missing data values and many more. So how then can we deal with these data anomalies? Data preprocessing concept therefore seeks to streamline and improve the quality of data hence making it more reliable. Data preprocessing does this by removing the extraneous information and mining the key features of the data to simplify the pattern detection process difficulties without disregarding any critical information.

1.1 Goals and objectives of data preprocessing

The major objectives of data preprocessing process are:
1. To reduce the size of the input space
2. To smoothen relationships
3. To normalize data
4. Reduce data outliers and
5. To extract data features

2. Literature Survey

Research shows that there are a number of data preprocessing techniques in data preprocessing namely: data cleaning, data integration, data transformation, data reduction and data discretization [3].

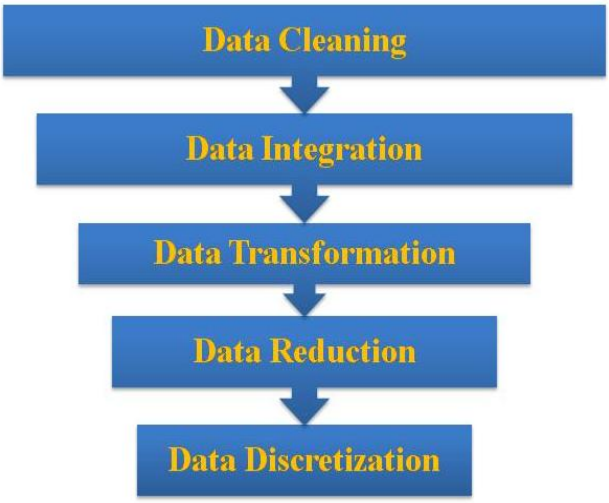

Figure 1: Showing data preprocessing techniques.

2.1 Data Cleaning

The data "cleaning" routine entails various tasks such as; data acquisition, filling missing data values, unifying date formats, conversion of nominal values to numeric data, identification of outliers and smoothening of noisy data, and correcting inconsistent data[1]. Dirty data however, is bound to causing confusion to data mining techniques hence it very necessary to clean mining data in order to reduce the size of the input space.

2.2 Data Integration

This phase deals with the combination of data from multiple sources like many databases, data cubes among others sources, into a coherent store. This procedure however, may happen in different forms as described below [1]:

1. Integration of schemas in order to incorporate metadata from different sources.
2. Identification of entities in order to identify the real world entities from multiple data sources that are directly related in terms of purpose, e.g., tbl.stuff_No.

3. Detection and resolution of data value conflicts. In some cases for instance, for the same real entity or attribute values from dissimilar sources are dissimilar for possible reasons like different representations or different scales, e.g., metric vs. British units

2.3 Data Transformation

Transformation process changes and merges the data into structural configurations that are appropriate for knowledge discovery. Research shows that transformation process achieves its objectives through the following sub-processes [2]:

1. Smoothing in order to remove outliers from the data
2. Aggregation procedure in order to provide the results summary or conducting the aggregation operations
3. Generalizing data values in so as to ensure that the low-level data is thereby replaced by higher-level concepts.
4. Data normalization by scaling the attribute data and setting them in a specified range.
5. Constructing the feature in order to construct new attributes and add them from a given set of attributes hence promoting better knowledge discovery process.

2.4 Data reduction

A data bank may contain millions of terabytes of data. This mass data brings about the problem of complexity in data analysis hence may take a very long time to run on the complete data set.[2]

Data reduction process is objectively done to obtain a reduced representation of the data set that is much smaller in size while strictly maintaining the integrity aspects of the original data. Data reduction process involves various strategies such as: [1] data cube aggregation, attribute subset selection, dimensionality reduction and discretization, and concept hierarchy generation.

1. Data cube aggregation – Data aggregation procedures are done to the data in the operation of a data cube.
2. Dimension reduction - Removal of irrelevant and redundant attributes.
3. Data compression - Deals with encoding mechanisms like data compression in order to reduce the data set size.
4. Numerosity reduction - The real data are replaced with smaller data representations such as parametric models that store only the model factors rather than the actual data.

2.5 Discretization and concept hierarchy generation

Here, the raw data values for continuous attributes are properly replaced with the created and divided ranges of attributes that are inform of intervals or higher conceptual levels [2]. With the concept hierarchies, mining of data at multiple levels are made feasible especially with the use interval labels that replace actual data values. E.g. The numeric values for the attribute age – at a higher level conceptual point of view, could be represented as: minor or adult.

3. Conclusion

Data preprocessing is quite fundamental because of the facts that real world data is of poor quality due to issues like presence of outliers, incompleteness or inconsistency problems. The major usage of data preprocessing technique is to ensure that there is data size reduction of the input space, smoothening of data relationships, data normalization, noise reduction and feature extraction(That is, unimportant data is ignored but at the same time maintaining the most important data characteristics). The commonly applied methods in data preprocessing are: data cleaning, data integration, data transformation, data reduction and data discretization.

4. Acknowledgement

I would like to thank the almighty God, my supervisor Dr. Wilson Cheruiyot, Emmilly and Antony Nzyoki for helping me to undertake this research from start to completion. May all of you be blessed so much.

References

[1] Jiawei Han et al,"Data mining, concept and techniques" .cs.sfu.ca, 2, Jan. 31, 2011. [Online]. Available: http://www.cs.sfu.ca

[2] Gregory Piatetsky-Shapiro et al,"Data mining concepts and preprocessing" kdnuggets.com , 2012. . [Online].Available: http://www.kdnuggets.com/data_mining_course/

[3] David L. Olson Dr., Dursun Delen Dr. "Advanced data mining Techniques "springer.com, 2008.[Online]. Available: http://www.springer.com/978-3-540-76916-3. [Accessed: Sept. 12, 2014].

[4] Tran, V.A., Hirose, O., Saethang, T., Nguyen, L.A.T., "D-IMPACT: A Data Preprocessing Algorithm to Improve the Performance of Clustering." Journal of Software Engineering and Applications, 7, 639-654. (2014)

Author Profile

Mathew Maingi has over a half a decade of experience in the field of computing and Information Technology. He specializes in database programming and database systems research and development. He is an expert in the area of developing strategies for innovation and creativity. He provides thought leadership and pursues strategies for engagements with the senior executives on Innovation in Business and Information Technology. He received the B.S. and M.S. degrees in Information Technology from Jomo Kenyatta University of Agriculture and Technology in 2011 and 2014, respectively. During 2011- 2014, he stayed in I.T. Research and lectured in various universities in Africa. He is now amongst the best and well recognized Information technology specialist not only in his home country Kenya, East Africa or the whole of Africa, but also globally.