Global Scientific JOURNALS

GSJ: Volume 7, Issue 10, October 2019, Online: ISSN 2320-9186

www.globalscientificjournal.com

Backward feature elimination and missing values ratio techniques for dimensionality reduction in data mining.

Isaac Museveni,Dr. Papias NIYIGENA

Department of Information Technology, University of Lay Adventist of Kigali

1.0 ABSTRACT

A data warehouse stores too much data from different sources and data mining is the process that we use to extract this data sets from the data warehouse. During this process however too much data that is unnecessary is extracted and this is what we are terming as large dimensionality rates. This research therefore is aiming at looking at various techniques that can be used to minimize this tendency of getting too much unnecessary data sets from the data warehouse. In this thesis we are going to use two techniques to reduce this tendency, the first technique is backward feature elimination where at a given iteration, the selected classification algorithm is trained on *n* input features, during this technique the process starts with all the columns and fields contained in the data sets,however it is at this point that it starts removing the least significant data sets including their features at each iteration hence improving the accuracy of the data.Another technique is missing values ratio which aims at removing columns with missing values.The use of these two techniques will not only reduce the dimensionality rates but also improve on algorithm speed and performance.The paper discusses a hybrid framework that combines the two algorithms with an aim of improving the quality of extracted data and speed at which data miners can interpret the information from the warehouse.

2.0 INTRODUCTION

In this era, the use of data warehouses has become so important due to the introduction of various industries dealing with data in their operations and these include both financial and manufacturing industries such as Banks and car manufacturing industry of which both have to

store large data amounts that they normally use in their operations. Therefore this implies that they need to use various data mining techniques to extract necessary data sets that they require, this is therefore very important to come up with techniques to help them extract necessary data sets leaving out what they do not need. Backward feature elimination and missing values ratio techniques would help them achieve this, these would help in data compressing and reducing the storage space required (Sunil Ray, July 2015).

3.0 LITERATURE REVIEW

3.1 Introduction

This chapter focuses on the existing theories and the literature on the work that has been done by other researchers. It will also focus on identifying the gaps in the existing solutions at the same time introducing the proposed framework and how it works in dimensionality reduction.

3.2 The existing frameworks

Due to increasing amount of data being used in many sectors today, there comes a great need for extracting this data from the DW. It is now possible to analyze large amounts of high-dimensional data with high-performance contemporary computers (Mizuta, M. 2012).Most researchers have looked at using one particular technique for dimensionality reduction. The highest reduction ratio without performance degradation is obtained by analyzing the decision cuts in many random forests (Random Forests/Ensemble Trees). However, even just counting the number of missing values, measuring the column variance, and measuring the correlation of pairs of columns can lead to a satisfactory reduction rate while keeping performance unaltered with respect to the baseline models (Rosaria Silipo, 2009).

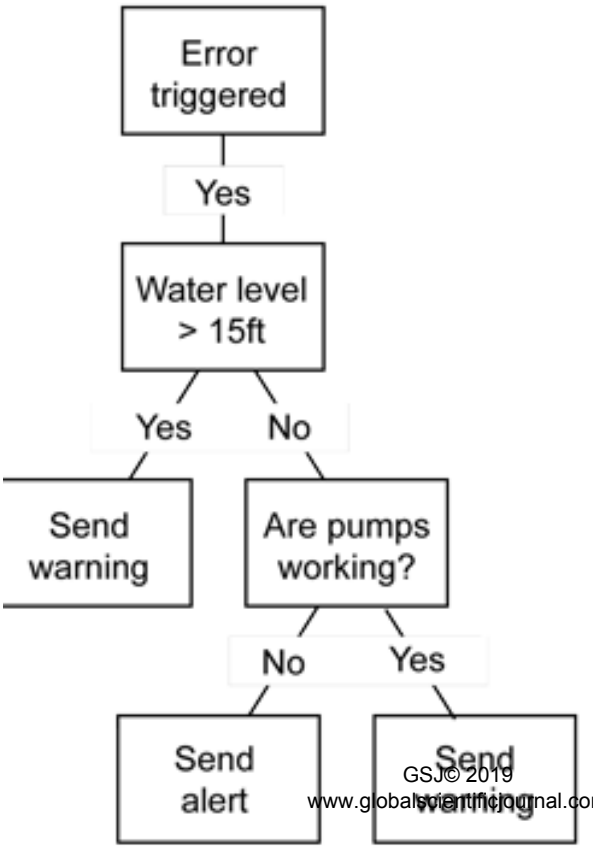
When using a single technique to reduce the data sets dimensions, one does not completely remove all redundant data sets.

3.3 Study of the existing framework for Dimensionality reduction

In this section, we are going to look at some of the proposed algorithms for dimensionality reduction. Techniques for dimensionality reduction in supervised or unsupervised learning tasks have attracted much attention in computer vision and pattern recognition. Among them, the linear algorithms Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been the two most popular because of their relative simplicity and effectiveness. Another linear technique called Locality Preserving Projections (LPP) has been proposed for dimensionality reduction that preserves local relationships within the data set and uncovers its essential manifold structure. (Yan Squishing, 2007).Another framework that was discussed by researchers is a Global Geometric Framework for Nonlinear Dimensionality Reduction (Tenenbaum, J. B., De Silva, V., & Langford, J. C. 2000).

4.0 Methodology

During this research in data mining, I used decision trees where it used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.



Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output.

4.1 Research Design

I used both qualitative and quantitative methods to carry out the research on different areas, as it will be shown on research areas and population. This included surveys, observations, focus groups and interviews.

4.2 Research Area

The research was taken from a group of people who deal with too much data sets in their places of work and that is in the Bank of Kigali where each district has got the DW for collecting some data from all the people in that district.

4.3 Data Collection

In this research, only secondary data was collected and it included the data and statistics from other researchers about data dimensionality reduction, and what other researchers discovered

about the existing algorithms. The secondary data was collected from modern research that has been done by other researchers. The sources included research papers, journals, books and any other online sources that provided relevant and up to date information that helped in the development of this framework. The sources were selected because of their rich up to date information about the dimensionality reduction.

5.0 Research results and discussion.

5.1 Introduction

The purpose of this study was to design a framework that combines two techniques which can used to reduce dimensionality in data mining. This chapter presents the research results and discussions on the performance of backward feature elimination and missing values ratio hybrid techniques.

5.2 Results Presentation

This was done by first identifying any financial or industrial firms that deal with a lot data and which may need extraction, the researcher opted for financial firm, a bank which owns a ware house with some number of distributed databases. In this bank, the researcher found out that they normally extract data on a 99% ratio daily which means the relevance and significance of the data extracted is very paramount.

The researcher put the bank on a test to extract a number of datasets of customers and check the relevance of data extracted, the bank accepted to extracted a sample of 30 customers' information, it was found out that the datasets produced was huge and irrelevant involving

columns that were empty and other similar values. The datasets were subjected to the two dimensionality reduction techniques as explained in chapter 3.

5.3 Data Extraction and comparison results

After extracting the 30 datasets of customers, each dataset was analyzed separately and the results recorded. For each group of data extracted was examined and the number of rows and columns produced were more than 100 and this involved some columns that did not have any data, it also involved similar values, for example 28 out of the 30 datasets extracted has two columns that give the same information such date of birth and age.27 out 30 datasets had 4-5 columns that were empty for example most of the datasets did not show the next of kin for various customers.

The results shown in the percentage form 0% to 100% were varying for each group of data. In this study, the customers' results were grouped into three levels depending on their dimensionality rates.

i. **Top level rate.** This was the highest level of dimensionality ratio with a high number of missing values and similar values, this was up to 75% loss of data.

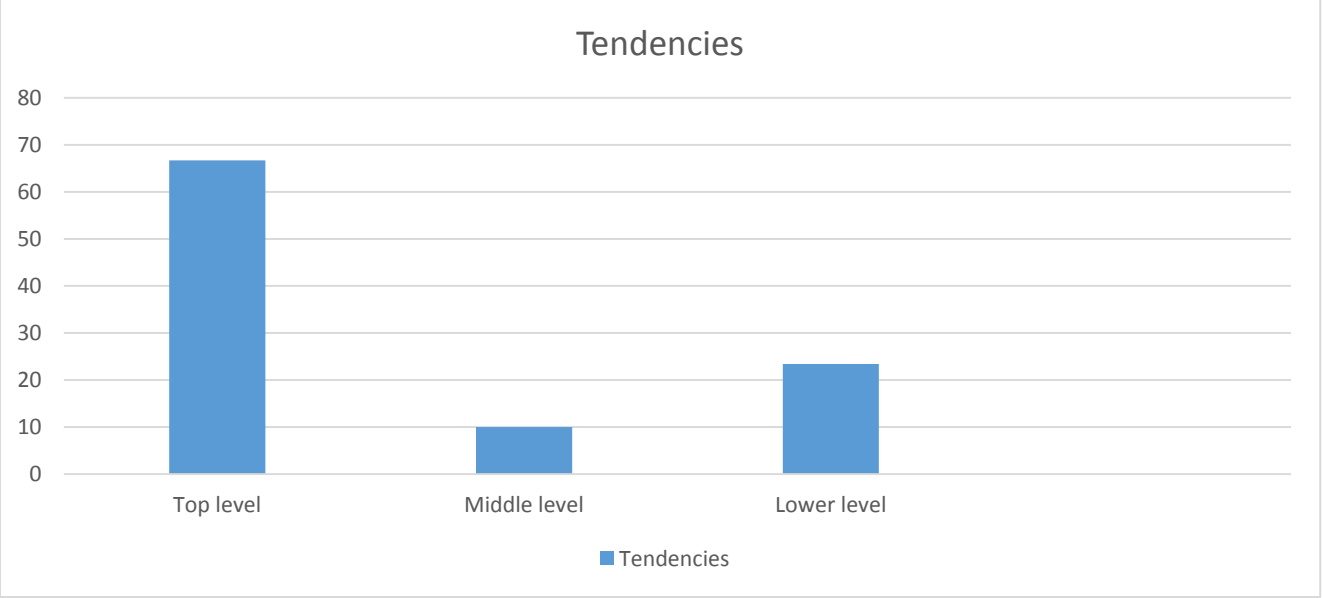
ii. **Middle level rate.** This was the middle level of dimensionality ratio with an average number of missing values and averagely similar values, this was up to 50% loss of data.

iii. **Lower level rate.** This was the lowest level of dimensionality ratio with the lowest number of missing values and similar values, this was up to at least 75% loss of data. This was considered the only level where the customers had less irregularities.

NO.	Levels	No. Of customers	Percentage	Remarks

1.	Top level	20	66.7%	This was the level with the biggest number of customers and yet it was having many irregularities.
2.	Middle level	3	10%	This was the level with the lowest number of customers and it had average irregularities.
3.	Lower level	7	23.2%	This was the level with fewer irregularities and compared to the number of the customers investigated, this was a small number and it brings a question of accuracy of the data extracted.

The graph below shows the percentages of different tendencies.



6.0 Acknowledgement

I would like acknowledge the work and assistance of my lecturer Prof. Wilson Cheruiyot who has tirelessly guided me and taught me how to prepare this thesis for the whole semester.

I wish to thank my classmates with whom I have discussed many issues in Information Technology that have been so important in writing this thesis.

I would like to recognize my family, my wife and parents for their tireless prayers and encouragement given to me.

7.0 References

lGuyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

(Tenenbaum, J. B., De Silva, V., & Langford, J. C. 2000).

H. Sato, Y. Hata, H. Masui, T. Tsumoto, J. Neurophysiol.55, 765 (1987).

M. E. Hasselmo, Behav. Brain Res. 67, 1 (1995).

M. G. Baxter, A. A. Chiba, Curr. Opin. Neurobiol. 9,178 (1999).

B. J. Everitt, T. W. Robbins, Annu. Rev. Psychol. 48,649 (1997).

R. Desimone, J. Duncan, Annu. Rev. Neurosci. 18, 193(1995).

P. C. Murphy, A. M. Sillito, Neuroscience 40, 13(1991).

M. Corbetta, F. M. Miezin, S. Dobmeyer, G. L. Shulman,S. E. Peterson, J. Neurosci. 11, 2383 (1991).

J. V. Haxby et al., J. Neurosci. 14, 6336 (1994).

A. Rosier, L. Cornette, G. A. Orban, Neuropsychobiology98 (1998).

M. E. Hasselmo, B. P. Wyble, G. V. Wallenstein, Hippocampus6, 69