DOI:10.26524/jms.2017.16

Study on retail marketing sale data using r software data cleaning and clustering algorithms

N.MARUDACHALAM, L.RAMESH
DR.AMBEDKAR GOVT.ARTS COLLEGE,
P.G & RESEARCH DEPARMENT OF COMPUTER SCIENCE,UNIVERSITY OF MADRAS

Abstract— Nowadays, data cleaning solutions are very essential for the large amount of data handling users in an industry and others. The data were collected from Retail Marketing sale data in terms of the mentioned attributes. Normally, data cleaning, deals with detecting, outlier detection, removing errors and inconsistencies from data in order to improve the quality of data. There are number of frameworks to handle the noisy data and inconsistencies in the market. While traditional data integration problems can deal with single data sources at instance level. The Hierarchical clusters and DBSCAN clusters were grouped with related similarities, analysis and Time taken to build model in different cluster mode was experimented using WEKA tool. It also focuses on different input retail marketing data by time calculated analysis. Clustering is one of the basic techniques often used in analyzing data sets. The Hierarchical and DBSCAN clustering Advantage and disadvantage also discussed.

Index Terms—Attributes, Data cleaning, Clustering

INTRODUCTION

Data cleaning is a process of identifying or determining expected problem when integrating data from different sources or from a single source. There are so many problems can be occurred in the data warehouse while loading or integrating data. The quality data can only be produced by cleaning the data and pre-processing it prior to loading it in the data warehouse. Data quality problems are present in single data collections, such as files and data bases, e.g., due to misspellings during data entry, missing information or other invalid data. This method gives the quality data for the end users or business people for of same kind data sources.

R is a system for statistical computation and graphics. It provides, among other things, a programming language, high level graphics, interfaces to other languages and debugging facilities.

It is important that an analyzes be carried out to identify some of the data cleaning techniques such that Identify the missing data Algorithm ,Outlier detection and inconsistent data of retail marketing data for using R Software. The comparison of Hierarchical clustering Algorithm and DBSCAN clustering Algorithm on the basis of response time.

II. PROCESS OF DATA CLEANING

1. Original Data: This data is Retail marketing sale data. It is last and previous sale data to taken in the experiments of this work.

2. Identify the Error Data: Which type of error to be occurred in the data set? To analyze the error and then rectify the error in the data. Types of error: Data entry problem, spelling mistake, Outlier data, incomplete data.

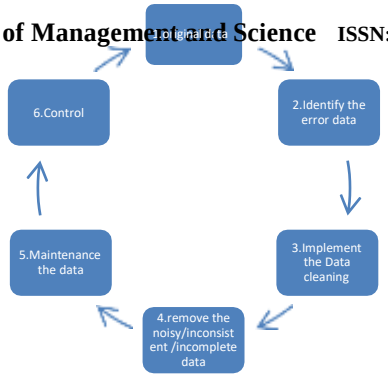
3. Implement Data Cleaning: Once the cleansing begins, it should begin to standardize and cleanse the flow of new data as it enters the system by creating scripts or workflows. These can be run in real-time or in batch (daily, weekly, monthly) depending on how much data has been taken for working. These routines can be applied to new data, or to previously keyed-in data.

4. Remove the noisy/inconsistent/incomplete using methods: Datacleaning ,Data methods,Data Analysis,Handling noisy data.

5. Maintenance the data: The database should be backed up continuously. The system should always be prepared for hardware or software failures and data loss. Procedures should be made as simple as possible to ensure that backups are regularly made. As the database involves with time and changes in information technology occur, data collection of data is essential to allow data access of formal data stored in former structure or design.

6.Control: We should be controlling your database on a whole. At the end, it is to bring the entire process full circle. Again revisit your plans from the first step and re-valuation. Make changes that outcome of the process and conduct regular reviews to make sure that the data cleaning is running with valuable and accuracy.

III. STEPS FOR PROCESS OF DATA CLEANING



IMPORTANCE OF DATA CLEANING

- Eliminate Errors
- Eliminate Redundancy
- Increase Data Reliability
- Deliver Accuracy
- Ensure Consistency
- Assure Completeness

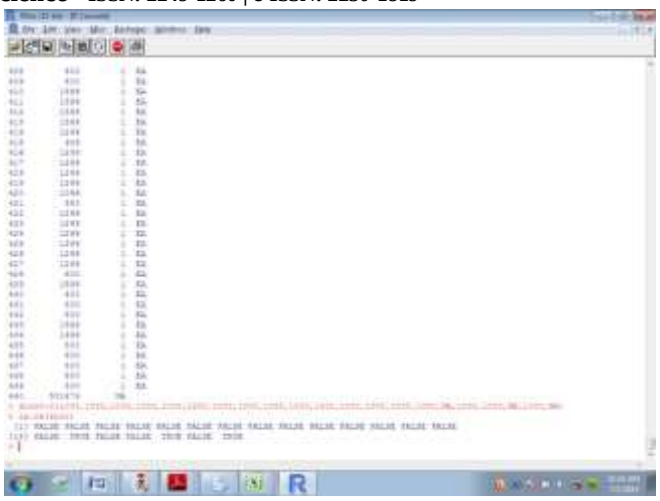
UPLOAD DATASET IN R SOFTWARE

DATA CLEANING USING R SOFTWARE

MISSING DATA

Missing data, or missing values, occur when no data value is stored for the variable in our point of view. Missing data are a common occurrence and can have a meaningful effect on the conclusions that can be written from the data .Missing data can occur because of non-responsibility, no information is delivered for several items or no information is provided for a whole unit. Some items are more sensitive for non-responsibility.

R software using a data cleaning by identifies the missing data.

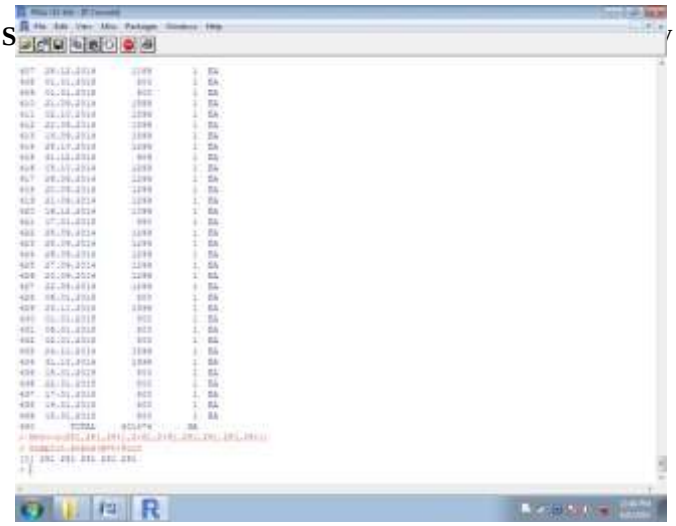


Identify Missing data using R Software commands:
>sloc<-c(1000,1000,1000,1000,NA,1000)
>is.na(sloc)
Output:[1]FALSE,FALSE,FALSE,FALSE,TRUE,FALSE

OUTLIER ANALYSIS USING R SOFTWARE

Outliers arise because of human error, instrument error, and natural deviations in populations, fraudulent behavior, and changes in behavior of systems or faults in systems. How the outlier detection system deals with the outlier depends on the application area. If the outlier indicates a typographical error by an entry clerk then the entry clerk can be notified and simply correct the error so the outlier will be restored to a normal record.

An outlier resulting from an instrument reading error can simply be expunged. A survey of human population features may include anomalies such as a handful of very tall people. Here the anomaly is purely natural, although the reading may be worth flagging for verification to ensure no errors, it should be included in the classification once it is verified.

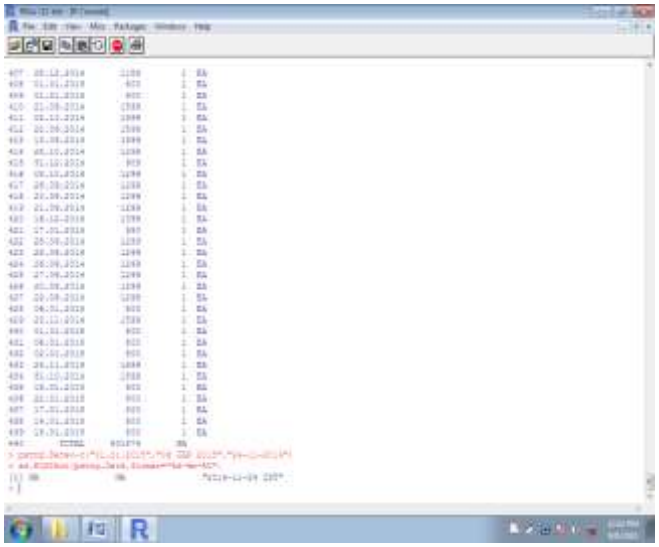


Identify Outlier Analysis Using R Software commands:
>Mvt<-c(251,251,2.51,25:1,251,251,25.1)
>boxplot.stats(Mvt)$out
Output:[1] 251 251 251 251 251 251 251 251 251 251

INCONSISTENT DATA USING R SOFTWARE

This occurs where two related databases do not use the same names lists, and when combined (or compared) show inconsistencies. For example, this may occur at botanic gardens between the Living Collection and the Herbarium; when merging databases of two specialists; or in museums between the collection database and the images database. Correcting involves checking one database against the other to identify the inconsistencies.

IDENTIFY MISSING DATA USING R SOFTWARE



Identify Inconsistent data using R Software commands:
>Pstng.Date<-c(–23.01.2015‖,‖06-jan-2015‖,‖24-11-2015‖)
>as.POSIXct(pstng.Date,format=‖%d-%m-%y‖)
Output:[1] NA NA 2014-11-2015

RESULT FOR DATA CLEANING USING R SOFTWARE

PARTICULAR	ORIGINAL DATA	CORRECT DATA ERROR DATA	MISSING DATA	OUTLIER DATA	INCONSISTENT DATA
Site	10000	470	200	270	-
Sloc	10000	970	500	120	350
Articledesc	10000	480	355	125	-
Mvt	10000	625	255	-	370
Mat.doc	10000	550	-	500	50
Item	10000	438	150	234	54
Posting date	10000	1079	900	56	123
Sales value	10000	745	255	400	90
Quantity	10000	980	210	-	770

HIERARCHICAL CLUSTERING ALGORITHM

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis.

DENSITY BASED CLUSTERING ALGORITHM

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm.Itusestheconceptof densityreachability and densityconnectivity. Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbors which are with in distance ε.

COMPARISON OF HIERARCHICAL CLUSTERING AND DENSITY BASED CLUSTERING ALGORITHM BASED ON CLUSTER MODE

1. Use training set
2. Supplied test set
3. Percentage split

Hierarchical Clustering & DBSCAN

10000 Attributes,10 Instances Based On Training Set

Algorithm	Use training set	Attributes	Instances
Hierarchical clustering	14.4 min	10000	10
DBSCAN Clustering	8.50 min	10000	10

Hierarchical Clustering & DBSCAN
10000 Attributes,10 Instances based on supplied Test

Algorithm	Use training set	Attributes	Instances
Hierarchical clustering	1.12 min	10000	10
DBSCAN Clustering	2.2 min	10000	10

iv. Comparison of hierarchical clustering and DBSCAN clustering using percentage split(instances:2000,Attributes:10)

Percentage split	Hierarchicl Clustering (in sec)	DBSCAN Clustering (in sec)	Instances	Attributes
10%	0.05 sec	0.02 sec	2000	10
40%	0.30 sec	0.25 sec	2000	10
50%	0.55 sec	0.45 sec	2000	10
70%	1.04 sec	0.55 sec	2000	10
90%	2.02 sec	1.35 sec	2000	10

Comparison of hierarchical clustering and DBSCAN clustering using percentage split (instances: 10000,Attributes:10)

Percentage split	Hierarchical Clustering (in sec)	DBSCAN Clustering (in sec)	Instances	Attributes
10%	0.05 sec	0.02 sec	2000	10
40%	0.30 sec	0.25 sec	2000	10
50%	0.55 sec	0.45 sec	2000	10
70%	1.04 sec	0.55 sec	2000	10
90%	2.02 sec	1.35 sec	2000	10

V. CONCLUSION

In this Work make use of data cleaning and data mining process in a Retail marketing sale database using Hierarchical and DBSCAN clustering methods to data cleaning usages and to analysis for clustering algorithm. The research information generated after the implementation of data cleaning and data mining technique may be helpful for Business people as well as for Human beings.

This work may improve data cleaning of retail marketing database; reduce missing data ratio by taking appropriate steps at right time to improve the quality of database for customer satisfaction. In the recent few years data cleaning and data mining techniques covers every area in our life. Data cleaning and mining techniques in mainly in the medical, banking,

insurances, Retail marketing, etc.

I have compared for Hierarchical clustering and DBSCAN clustering Algorithm of basis in time taken to build model for in performing seconds ,To analyzed was DBSCAN clustering Algorithm is the best one in this experiment.

VI. REFERENCES

1. ERCreator software at http://www.modelcreator.com.
2. Kettle software at http://www.kettle.be/index.htm.
3. Martinez, A. and Hammer, J. Making quality count in biological data sources. In Proceedings of the International
4. Workshop on Information Quality in Information Systems (IQIS 2005).
5. Pipino, L., Yang, W. Lee, Y., Wang, R. Data quality assessment. Communications of ACM. 45, 4 (April 2002).
6. Preen, A., Marsh, H., Lawler I.R., Prince, R.I.T., ShepherdR. 1997 Distribution and abundance of dugongs, turtles,dolphins and other Megafauna in Shark Bay, Ningaloo Reef and Exmouth Gulf, Australia Wildlife Research. 24(1997).
7. Raman, V., Do, H. Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.
8. Redman, T. The impact of poor data quality on the typical enterprise. Communications of ACM. 41, 2 (February 1998).
9. Scannapieco, M., Virgillito, A., Marchetti, M., Mecella, M.,Baldoni, R. The DaQuinCIS Architecture: a platform for exchanging and improving data quality in cooperative information systems, Information Systems, 29 2(2004).
10. Feinerer. An introduction to text mining in R. R News, 8(2):19{22, Oct. 2008. URL http://CRAN.R-project. org/doc/Rnews/.