International Journal of Research in Advent Technology, Vol.2, No.9, September 2014
E-ISSN: 2321-9637

A Study on Data Cleansing and Classification Algorithms for Large Dataset Systems

R. Deepa[1],Dr. R Manicka Chezian[2],
Ph.D Scholar, Department of computer Science (Aided), NGM College Pollachi, E Mail: phd.deepakarthikeyan@gmail.com[1]
Associate professor, Department of Computer Science (Aided), Pollachi [2]

Abstract:-In today's competitive environment, there is a need for more precise information for a better decision making. Yet the inconsistency in the data submitted makes it difficult to aggregate data and analyze results which may lead to delay or data compromises in the reporting of results. This paper presents a study of the different algorithms available to clean very large dataset to get for the need for more consistent data. The data cleaning algorithms can increase the quality and also it reduces the computational cost after finding and removing the outliers

Index Terms: Dataset, Clean, Computational Cost, outliers.

1. INTRODUCTION

Data mining is the process of database analysis that attempts to discover useful information's from a large dataset. The analysis uses several advanced statistical and other methods, like cluster analysis, and sometimes it uses artificial intelligence or neural network techniques.

A major objective of data mining is to find previously hidden relationships between the data, especially when the data is collected from different databases. Data mining is used in several areas like insurance, banking, retail, astronomy, medicine detection of criminals and terrorists. The process of converting data to knowledge has several phases that is shown in the figure -1

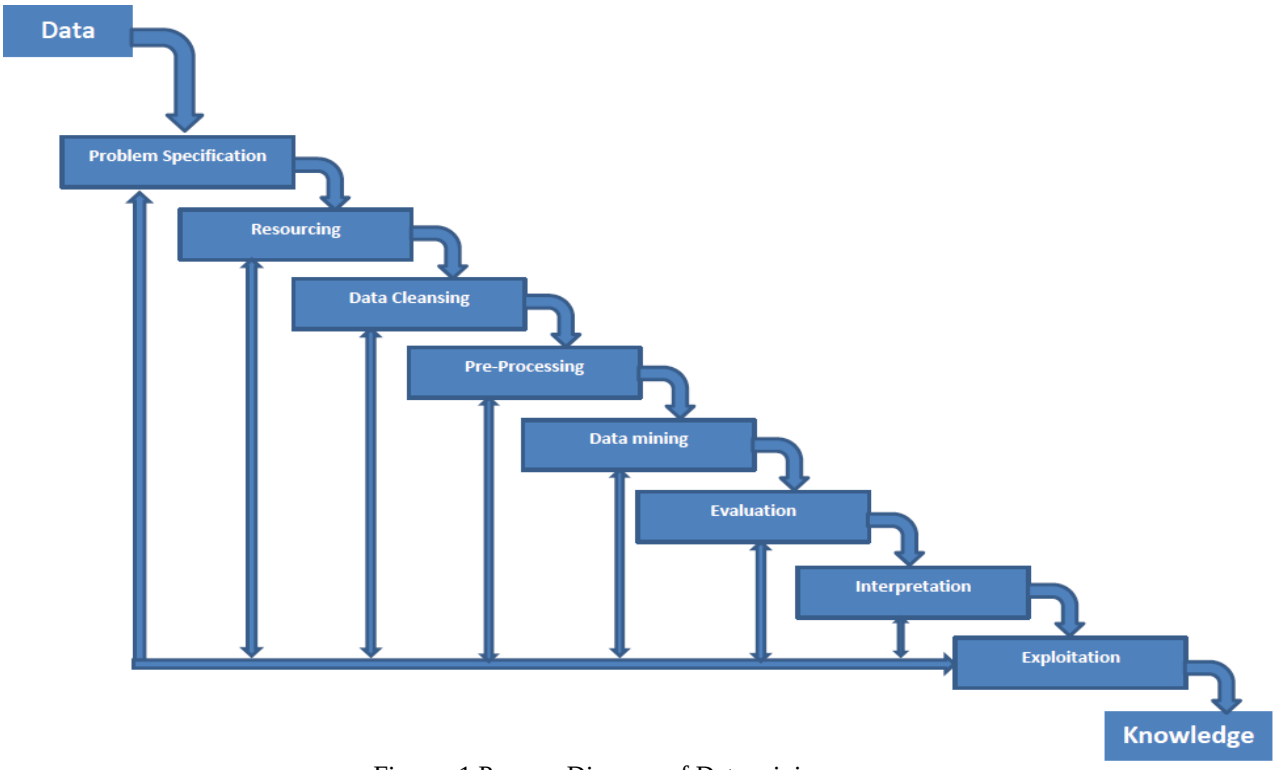

Figure -1 Process Diagram of Data mining

2. DATA CLEANSING

Data cleansing, data cleaning or data scrubbing is the process of finding and rectifying improper or wrong information or records from a large record set, collection of records (table), or from collection of tables (database).this concept is mainly is used in databases, the term data cleansing refers to finding incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and subject to replacing, modifying, or deleting this outlier data or coarse data. After data cleansing, the data set will be consistent for analyzing or any other operations with other similar data sets in the system. The inconsistencies data should be analyzed and detected or removed from the original database. The process diagram is shown for the data cleansing is shown in the figure-2.

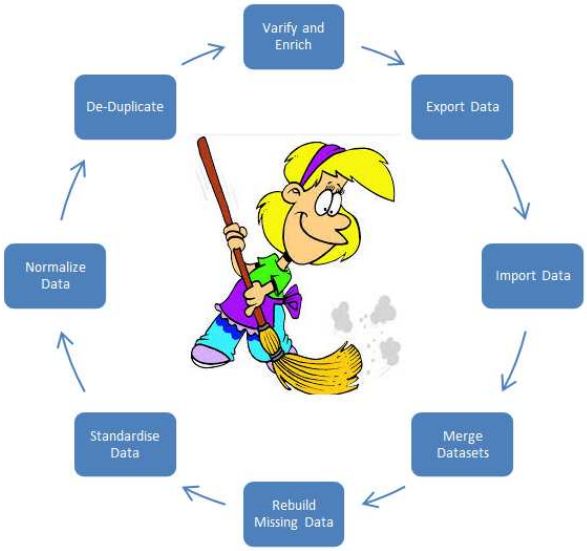

Figure -2 Data Cleansing

3. DATA CLEANING APPROACHES

In general, data cleaning has several phases

Data analysis: the first phase of data cleansing is data analyses because the data is collected from the heterogeneous background or from different data sets or databases. So the possibility of errors or bugs is high In order to detect and remove such errors and inconsistencies in the data set a detailed data analysis is required. In addition the analyzed data must be subjected to manual verification or inspection with the data or data samples and some sort of analysis programs should be used to gain metadata about the data properties and find the data quality problems.

Definition of transformation workflow and mapping rules: this phase is purely depends upon the number of data sources. To fulfill this phase large number of transformation or cleaning steps may have to be involved. This is purely based on their degree of heterogeneity and the "dirtyness" of the data. Sometime, a schema translation is used to chart sources to a common data model. Typically a relational representation is used for data ware houses. Early data cleaning steps can correct single-source instance problems and prepare the data for integration. Later steps deal with schema/data integration and cleaning multi-source instance problems, The schema-related data transformations as well as the cleaning steps should be specified by a declarative high query and mapping language as much as possible, to enable automatic generation of the transformation code. In addition, it should be possible to embed the user-written cleaning code and special purpose tools during a data transformation workflow. The transformation steps may request user feedback on data instances for which they have no built-in cleaning logic.

Verification: The correctness and effectiveness of a transformation workflow and the transformation definitions should be tested and evaluated both manually and algorithmically, e.g., on a sample or copy of the source data, to improve the definitions if necessary. It requires Multiple iterations of the analysis in order to improve quality of the iterated or analyzed dataset, in addition design and verification steps may be needed, e.g., since some errors found onlyafter applying some kind of transformations techniques.

Transformation: Execution of the transformation steps is maintained by either running the ETL workflow for loading and refreshing a data warehouse or when answering queries on datasets or multiple tables.

Backflow of cleaned data: After each error are removed from the dataset , the cleaned data should also replace the outlier or dirty or unwanted data in the original sources in order to give legacy applications the improved data and to avoid repetition of cleaning work for future data extractions. For data warehousing, the cleaned data is available from the data staging area.

4. EMPIRICAL REVIEW

Most of the data cleaning research deals with schema translation and schema integration, data cleaning has received only little attention in the

research community, but effective data cleaning methods may improve the effectiveness of data mining methods or the accuracy of any data mining techniques.. Many researchers and their research work are focused on the problem of data cleaning and they suggest many methodologies, optimization techniques and variety of new algorithms have been proposed different algorithms to clean outlier data.

An effective data cleaning method is proposed byWejje Wei, Mingwei Zhang, Bin Zhang Xiaochun Tang. The authors have introduced a new methodology based on association rule mining method. This association rule mining method exploits the business rules provided by the association rules mined from multiple data sources and this methodology creates promising business rules for individual data sources. The association rules mining methodology is achieved and reduced computational cost and accuracy in cleaning is reported[1].

Applied Brain and Vision Science-Data cleaning algorithm is a novel algorithm which is designed for cleaning the EEG resulting from the brain function of "stationary" behavior such as for an eyes-open or eyes-closed data collection paradigm. This algorithm assumes that the useful information contained in the EEG data is stationary. That is, it assumes that there is very few or small change in the on-going statistic of the signals of interest contained in the EEG data. the experimental results shows that this algorithm gives good results when removing momentary artifacts in EEG collected on both condition such as eyes-closed or eyes opened state[2].

Chaudhuri, S., Dayal have proposed a new execution model and novel algorithms . In this approach users can involve and suggest data cleaning requirements and specifications declaratively which helps to perform the cleaning efficiently. This methodology uses a dummy a set of bibliographic references used to construct the Citesser Web Site. Also the researchers have proposed a model for cleaning textual records, this results that meaningful information can be processed from the cleaned data set [4].

YiqunLiu,MinZhang,LiyunRu,Shaoping Ma[3] has proposed a novel learning-based algorithm. This finds very good results with reducing the web pages. This is done by calculating user need. The methodology gives only the most matching websites and other sites are treated as outliers. The results show that how the retrieval target pages can be separated from minimal quality pages using query-independent features embedded in the cleansing algorithms.

A threshold based data cleaning method is proposed by Timothy Ohanekwu, C.I.Ezeife [5]. This algorithm uses a technique which eliminates the need to rely on match threshold, which is achieved by defining smart tokens. So the threshold calculation is used for identifying outliers available in the dataset. This methodology also eliminates the need to use the entire long string records which needs multiple passes, for identifying the outliers.

Chris Mayfield et. al [7] has used a statistical method for integrated data cleaning and imputation. They focus on exploiting the statistical relationships between records in database; this methodology has a new approach to analyze the statistical dependencies between the records. The methodology automatically calculates these dependencies and also it has a capability to fill in missing values at the same time as detecting and correcting faults.

Yu Qian, Kang Zhang[6] has expressed a promising issue in the use of picturing for data mining: choosing proper parameters in the area of spatial data cleaning algorithms. This leads to improve the performance of cleansing algorithm through its visualization. The other benefit is its poetries and characteristics of algorithm and its features of dataset where visualized as feedback to the user.

A novel Data cleaning algorithm purely based on mathematical morphology is proposed by S.Tang in the area of bioinformatics and medical image understanding. In this area the possibility of frequent noise that occurs and deteriorates the performance are classified; This scenario leads to improve the mechanism of data cleansing in the training data. Also they deal with data noise is firstly revived and this noise cleansing mathematical morphology is used.

Kazi Shah Nawaz Ripon et. Al [9]. Proposed novel methodology for identifying and cleansing the identical tuples in a dataset. The algorithm is checked against the computational cost and they proved that this required less computational cost. Also they proposed the enhanced version of significant transitive rule.

PayalPahwa, Rajiv Arora, Garima Thakur has proposed addresses issues which related to identification and rectification of outlier which is available in the dataset. They have conducted an effective survey on existing data cleansing techniques and also discussed the major problems in the existing

methods. They proposed new architecture to overcome the limitation of existing methodology.

SurbhiAnand and Rinkle Rani Aggarwal has proposed[10] addresses issues which related to identification and rectification of outlier which is available in the dataset using the extension of the files. Also, it analyses various factors like data quality. They have conducted an effective survey on existing data cleansing techniques and also discussed the major problems in the existing methods. They proposed new architecture to overcome the limitation of existing methodology.

Two novel data cleansing algorithms are discussed and proposed by R.Kavitakumar, Dr. RM. Chandrasekaran the algorithm effectively implemented using famous data mining techniques to detect and correct the attribute without external reference. First one is Context-dependent attribute based detection and correction and second one is Context-independent attribute based detection and correction.

JieGupreposed a random forest based technique[12] and sampling methods to identify the potential buyers. This method has two phases: data cleaning and classification, both methods is purely based on random forest.

SubhiAnand, RinkleRani[14] is preposed a novel methodology to cleanse World Wide Web is a monolithic repository. They emphasize on the Web Usage and content Mining process and exploits in the area of data cleaning.

Li Zhao, Sung Sam Yuan, Sun Peng and Ling Tok Wang They propose [15] a method based on based on the longest common subsequence. This method called LCSS and explained the possesses with its desired properties. This paper also includes two novel detection and correction methods, SNM-IN and SNM-INOUT, which are the optimized or enhanced version of detection and correction method called SNM.

Aye T.T[16] has explained the data cleaning algorithm eliminates inconsistent or unwanted or dirty items in the preprocessed data.

Yan Cai-rong, Sun Gui-ning, Gao Nian-gao[17] has analyzed the limitation of traditional methods in knowledge base, an extended tree-like knowledge base and proposed a novel knowledge base data cleaning algorithm. The new method finds atomic value in between selected nodes firstly, after finding the value analyses started in order to their relations, based on the analysis it deletes the outlier objects, and calculates and stored an atomic value sequence based on weights is assigned.

ERACER method is proposed Chris Mayfield, Jennifer Neville, Sunil Prabhakar[18]. This paper presents framework using iterative statistical method for detecting missing information and correcting such bugs automatically. Belief propagation and relational dependency networks methods are used, and this method includes approximate inference algorithm.

Mohammad, H.H. is developed a method for identifying dirty data. Parsing algorithm is used to identify of outlier data. So this method effectively used K-nearest Neighbour algorithm for the finding the outlier and cleaning the data.

Shawn R. Jeffery, Minos Garofalakis, Michel J. Franklin [19] has proposed SMURF method, this is the first declarative, adaptive smoothing filter algorithm for effective RFID data cleaning.

Manuel CastejonLimas, et. al [20] implemented a new method for dirty data identification and data correction for both normal and no normal multivariate dataset. The new method is based on an iterated local fit. The method implemented without priori metric calculations. This approach supported by finite mixture clustering which leads to achieve best results using large data sets.

KollayutKaewbuadee,YaowadeeTemtanapat has developed an outlier detection engine by combining an FD discovery technique with an existing outlier detection technique and this optimization called “Selective Value”. This leads to decrease the number of identified FDs.

5. CONCLUSION

This paper surveys the data cleansing algorithm for large dataset and some fundamental steps in the data cleaning and data mining. Data cleaning is a very is very young field in the area of computer science research. This paper represents the current research and practices in data cleansing for large data set.

Although the large number of tools indicates both the importance and difficulty of the cleaning problem. This paper discussed several implementation of various algorithm effectively used in data cleaning which deserve for further research.

REFERENCES

[1] Weijie Wei, Mingwei Zhang, Bin Zhang "A Data Cleaning Method Based on Association Rules" , Northeastern University, Shenyang PP.1-6

[2] Applied Brain and Vision Science-Data cleaning algorithm

[3] Yiqun Liu, Min Zhang, Liyun Ru, Shaoping, "Data Cleansing for Web Information Retrieval using Query Independent Features" Journal of the American society for information science and technology, Vol58(12) pp-1-15

[4] Chaudhuri, S., Dayal, U" An Overview of Data Warehousing and OLAP Technology".ACM SIGMOD Record 26(1), 1997.

[5] Timothy E. Ohanekwu and C.I. Ezeife ,"A Token-Based Data Cleaning Technique for Data Warehouse systems" -University of Windsor P.P. 7-12

[6] Yu Qian, Kang Zhang ,"The role of visualization in effective data cleaning" SAC '05 Proceedings of the 2005 ACM symposium on Applied computing P.P 1239-1243

[7] Chris Mayfield, Jennifer Neville, Sunil Prabahakar, "A Statistical Method for Integrating Data Cleaning and Imputation" - Purdue University(Computer Science report-2009) Report Number -09 -008

[8] Sheng Tang "Data cleansing based on mathematical morphology" published in ICBBE 2008 the second InternationalConference-2008

[9] Kazi Shah Nawaz Ripon, Ashiqur Rahman and G.M. AtiqurRahaman"A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates",Journal Of Computers, VOL. 5, NO. 12, December 2010 . P.P 1800-1809

[10] SurbhiAnand and Rinkle Rani Aggarwal "An Efficient Algorithm for Data Cleaning of Log File using File Extensions" International Journal of Computer Applications Volume 48– No.8, June 2012. P.P 13-18

[11] R. Kavitha Kumar and Dr. R.M Chandrasekaran"Attribute correction-data cleaning using association rule and clustering methods" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.2, March 2011. P.P 22 - 32

[12]JieGu - Random Forest Based Imbalanced Data Cleaning and Classification – JieGu- Software School of Tsinghua University, China .P.P.1-7

[13] Sheng TANG and Si-ping CHEN "Data Cleansing Based on Mathematic Morphology" Bioinformatics and Biomedical Engineering, 2008.ICBBE 2008. The 2nd International Conference on May(IEEE-EXplore) 2008 755 - 758

[14] SurabhiAnand ,Rinkle Rani Aggarwal. "An efficient Algorithm for Data Cleaning of Log File using FileExtension" International journal of Computer Appliactions June-2012 Vol - 48(8).PP 13-18,

[15] Li Zhao, Sung Sam Yang, Sum Peng and Ling Tock Wang " A New Efficent Data Clencing Method" Springer - DEXA 2002 P.P 484 -804

[16] Aye, T.T. "Web log cleaning for mining of web usage patterns" Computer Research and Development (ICCRD), 2011 3rd International Conference(IEEE Expore) Vol-2 P.P 490 - 494

[17] Yan Cai-rong,SUNGui-ning , GAO Nian-gao ,"Mass Data Cleaning Algorithm based on extended tree-likeknowledge base" –Computer Enginerring and application –PP-146-148

[18] Chris Mayfield, Jennifer Neville and Sunil Prabhakar "ERACER-A database approach for statistical inference and datacleaning" SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA

[19] Shawn R. Jeffery, Minos Garofalakis and Michael J. Franklin "Adaptive Cleaning for RFID Data Streams" ACM VLDB '06, September 1215, 2006, Seoul, Korea. P.P 163-174

[20] Manuel CastejonLimas, Joaquin B. Ordieres Mere, Francisco J. Martinezn de Pison ,Ascacibar and Eliseo P. Vergara Gonzaalez"Outlier Detection and Data Cleaning in Multivariate Non-NormalSamples: The PAELLA Algorithm" Data Mining and Knowledge Discovery,Vol 9, P.P 171–187

[21] Mohamed H.H "E-Clean A Data Cleaning" Informatics and Computational Intelligence (ICI) 2011 (IEEE Explore)

BIOGRAPHY

R.Deepa received her Bsc.Statistics from P.S.G College of Arts and Science College, Coimbatore, India. She had her Master of Computer Applications from Bharathidasan University, Trichy, India. She holds MPhil, in Computer Science from Bharathiar University, Coimbatore, India. She has 9 years of experience in teaching. She is presently working as an Assistant Professor in NGM College, Pollachi. Her research interest includes Data Mining, Big Data Management, and Image Compression. Now she is pursuing her Ph.D Computer Science in Dr. Mahalingam Center for Research and Development at NGM College Pollachi.

Dr. R.Manicka chezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published thirty papers in international/national journal and conferences: He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.