© 2016 IJSRSET | Volume 2 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

An Enhanced Common Data Cleaning Framework for Data Mining

Agusthiyar. R[1], Dr. K. Narashiman[2]

[1]Research Scholar, Anna University, Chennai, Tamil Nadu India
[2]Professor & Director, AUTVS Center for Quality Management, Anna University, Chennai, Tamil Nadu India

ABSTRACT

In this information era, there is a huge availability of data but the information is not enough to meet the requirements. This creates an urgent need for data cleaning and data cleaning solutions become highly important for data mining users. Normally, data cleaning deals with detecting, eliminating errors and inconsistencies in large data sets. For any real world data set, doing this task manually is very cumbersome as it involves huge amount of human resource and time. This means several organizations spend millions of dollars per year to detect data errors. Due to this wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered one of the biggest problems in data warehousing. Normally the data cleaning is required when multiple data sources need to be integrated. In this research work, an Enhanced Common Data Cleaning (ECDC) framework has been developed and proposed.

Keywords: Data cleaning, Data mining, Extract, Transform and Load (ETL), Extensible Markup Language (XML), Enhanced Common Data Cleaning (ECDC)

I. INTRODUCTION

Recent days, data availability is abundant but the information becomes meaningless if it is not arranged or structured in a proper manner. Examples of the data mining applications like customer relationship management, telecommunication, retail management, weather forecasting, financial analysis, medical analysis and health care systems require quality data for processing their information. If the data has any one of the following issues like misspelling, inconsistency, irrelevance, inaccuracy and timeliness then the decision making process will fail. This will lead the organization to trouble. So, the data cleaning solutions are very essential in such circumstances. In this regard an Enhanced Common Data Cleaning (ECDC) framework has been developed and proposed. It has three phases, namely:-

i) Schema level data cleaning using Extensible Markup language (XML) constraints (SumonShahriar&SarawatAnam 2009)
ii) Extract Transform and Load (ETL) based data cleaning with smart tokens (Erhard Rahm &Hong-Hai Do 2000, Ezeife, CI & Timothy E Ohanekwu 2005)
iii) Data cleaning using Neural Network (Wei Wei 2001)

ECDC framework can be used to clean the data effortlessly and give the cleansed data for good decision making in business organizations and large data handling users.

The customer detail of any organization is a very important source or asset. If the company has branches in different places then the data source will be integrated in one place, at that time data quality problem in schema level constraints will arise. The data quality issues in the schema level could be rectified by the modified XML constraints of the ECDC framework. If the dataset has the noisy values, it will affect all the transaction of the organization because the data source is directly proportional to the organization's growth and decision support system. There is another scenario. Students residing in rural areas of Tamil Nadu get their Voter ID from their respective assembly constituency. After their

Graduation or Post Graduation studies, students migrate to metropolitan cities owing to varied job opportunities. If the student decides to get a new Voter's ID from the constituency where he/she resides, in this scenario, there is a duplication of Voter's ID. This issue will be resolved using ETL based data cleaning with smart tokens of ECDC framework.

II. METHODS AND MATERIAL

1. Literature Study

Due to the wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered to be one of the biggest problems in data warehousing. Data cleaning is the process of cleaning irreverent data and inconsistencies in large datasets. Normally the data cleaning is required when multiple data sources are integrated. This research work investigates current data cleaning methods, approaches, data quality oriented data mining and data warehousing frameworks and designs in the previous years. From this literature study, there is lacking in the standard data cleaning procedure for the researchers, and there is a survey about the data (field) error rates in the data acquisition phase which in typically around 5% or even more using sophisticated measures for error prevention. Recent studies have shown that 40% of the collected data is dirty in one way or another. During data cleaning, multiple records representing the same real life object are identified, assigned only one unique database identification, and only one copy of exact duplicate records is retained. So, this research work mainly concentrates on data cleaning in large data sets. In this view, the literature study about this domain is made in the following journal papers, articles and conferences. The literature survey was made for the period of four decades, to mention a few: Five journals from 1980 to 1989, 31 journals from 1990 to 1999, 39 journals from 2000 to 2005, 29 journals from 2010 to 2015 and in addition 12 conferences. From the literature study, it was ascertained that the data cleaning is very essential for data mining in large data sets.

As a conventional procedure, the data cleaning solutions can only be managed with data mining functions like Minimum, Average, and Maximum. The Minimum, Average, and Maximum functions are used to replace the noise data like missing and irrelevant values. In light of the above shortcomings and limitations there is an urgent need for common data cleaning framework. In continuation with this research interest, it is proposed to create an ECDC framework for data quality problem of single source and multi-source problems in schema level and instance level. Further the researcher proposed an ECDC framework in this research work. It has three phases of processes to do the data cleaning in large data sets. This was confirmed by the experiments done by the researcher for ten data sets such as Iris data set, Voter's data set, Ionosphere data set, Cylinder band data set, Mushroom data set, Adult data set, Cloud data set, Credit-approval data set, NLBB salary data set and Hospital data set.

2. Methodology

This research methodology section explains how the common data cleaning framework is organized. In this regard, an ECDC framework is proposed and it has three phases viz., i) Schema level data cleaning using XML constraints, ii) ETL based data cleaning with smart tokens and iii) Data cleaning using Neural Network. As the conventional procedure, the data cleaning solutions can only be managed with data mining functions like Minimum, Average, and Maximum. But in this work, the ECDC framework is proposed and tested with 6 data sets. The result has improved significantly. The first two phases were tested with clustering technique for k means algorithm in Rapid Miner tool. Both the results were improved and the same is under evidenced result and discussion title (4). The third phase, Data cleaning using Neural Network was tested in the Mat Lab tool. This tool has nntool, nftool, and nctool to bring, fitting, and create the neural networks appropriately. Initially the neural network is created with input, output, and error and the input is trained by the algorithm, if the desired output was not met. After the training, the neural network was trained and finds the noise values position. Then it was replaced by the array values given in the algorithm.

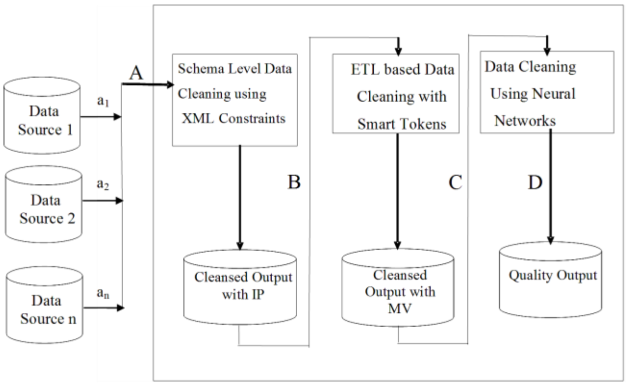
A
Data Source 1
a_1
Data Source 2
a_2
Data Source n
a_n
Schema Level Data Cleaning using XML Constraints
ETL based Data Cleaning with Smart Tokens
Data Cleaning Using Neural Networks
B
C
D
Cleansed Output with IP
Cleansed Output with MV
Quality Output

Figure1. An ECDC Framework for Data Cleaning

In ECDC Framework for Data Cleaning shown in Figure 1, A denotes the sum of a_1, a_2, and a_n the noisy data sets, (1, 2, and n) which are going to be cleaned. In the first phase, the Schema level issues will be rectified using modified XML constraints and the data is stored in the database. But, it may contain instance level problems, that need further cleaning. Then the data will be sent to the second phase, where B denotes the cleansed data with Instance Problems (IP). In this phase, instance problem will be rectified with smart tokens. C denotes the cleansed data but may contain Missing Values (MV). This data will be sent to the third phase. In this phase, data will be cleaned using Neural Network whereas it will detect and eliminate the missing values. Finally, D the quality output is stored in the database.

ECDC framework has three phases such as i) Schema level data cleaning using XML constraints, ii) ETL based data cleaning with smart tokens and iii) Data cleaning using Neural Network. Initially the data set is loaded into this framework .In this framework, the schema level issues will be checked in the first phase and if the data set has schema level issues like integrity constraints, poor schema designs and uniqueness problems it will be rectified using modified XML constraints. In this phase, there are five levels of process namely A, B, C, D and E were used to solve the schema level issues. A-selecting and identifying the noisy data from the data set. B-applying preprocess technique to segregate the data, C- applying modified XML constraints to redefine the schema, D- post processing for noise removal and E-cleansed data with instance level problems.

In the second phase, an ETL based data cleaning with smart tokens will be used to solve the instance level issues like misspelling, redundancy and contradictory values, inconsistent aggregating and inconsistent timings. In this phase, there are eight steps to resolve the issues. First two steps were used to select and combine the attributes of two different sources. In third step, choosing the unique attribute, it will act as the smart token of the table. Fourth step used to compare the fields of two tables, whether it is same or different. Fifth step detecting the noise and duplicate values based on the smart tokens. Sixth step filtering and eliminating the noise and duplicate values. In the Seventh step clustering algorithm is applying for segregating the data. In the eighth step, cleansed data will be stored into the final data warehouse. This framework has the following four stages of processes.

In the third phase, Data Cleaning using Neural Networks will be used to detecting and eliminating the missing values. It has four steps. In first get the data set of noise values. Secondly, FFNN method was applied into this data set for training the prediction array of inputs. Thirdly, BPNN are used to find the prediction of the value and location of the missing values. Finally, Execute the SQL query to replace the missing values according to the findings. In this work, the messy data have been applied in the proposed algorithm and verified the performance of the algorithm on missing and inconsistent data of data source. It has improved the accuracy and completeness of the dataset in a better way.

III. RESULTS AND DISCUSSION

A In this research the Iris data set, Voter's data set, Ionosphere data set, Cylinder band data set, Mushroom data set, Adult data set, Cloud data set, Credit-approval data set, NLBB salary data set and Hospital data set of UCI repository were tested for the proposed ECDC framework. These data sets were tested in all the three phases of this framework.

Table 1. List of Data sets used for testing

S. No	Data Set	No. of Attributes	No. of Numeric Instances	No. of Categorical Instances	No. of Mixed Instances	Missing Values
1.	Iris Data Set	5	150	0	0	No
2.	Voter's Data Set	34	0	435	0	Yes
3.	Ionosphere Data Set	34	351	0	0	Yes
4.	Cylinder bands Data Set	39	0	0	512	Yes
5.	Mushroom Data Set	22	0	8124	0	Yes
6.	Adult Data Set	15	0	0	32561	Yes
7.	Cloud Data Set	10	1024	0	0	No
8.	Credit-approval Data Set	15	0	0	690	Yes
9.	NLBB Data Set	4	0	438	0	Yes
10.	Hospital Data Set	4	20	0	0	Yes

The performance of the ECDC framework was tested on the data sets listed in the Table 4.2. The performance of the data mining algorithm was calculated by the following parameters: sensitivity, specificity, accuracy, precision, recall and F- measure. All the datasets were applied in the clustering methods of K-means, fcm and mode accordingly. The numeric data sets were applied in the k-means and fcm methods of clustering and the categorical data set was applied in the mode method of clustering. From the clustered output Tp (True positive, Tn (True negative), Fp (False positive) and Fn (False negative) values are measured and sensitivity, specificity, accuracy, precision, recall and F-measure values are calculated.

Table 2. Performance measurement of the data mining algorithm (K-means, mode and fcm)

S. No	Data Set	Sensitivity	Specificity	Accuracy	Precision	Recall	F Measure
1	Iris Data Set	80	75	77.7778	0.8000	4	3.2000
2	Voter's Data Set	90.9091	33.3333	89.3805	0.9804	20	19.6078
3	Ionosphere Data Set	70.5882	66.6667	69.1358	0.7826	3.6000	2.8174
4	Cylinder bands Data Set	92.0245	28.5714	89.4118	0.9677	18.750	18.1452
5	Mushroom Data Set	76.9231	66.6667	73.1707	0.8000	3.5714	2.8571
6	Adult Data Set	93.7500	33.3333	84.2105	0.8824	7.5000	6.6176
7	Cloud Data Set	85.7143	40	76.9231	0.8571	4.5000	3.8571
8	Credit-approval Data Set	57.1429	40	50	0.5714	1	0.5714
9	NLBB salaries Data Set	67	64	78.8	0.8	5.5	3.3342
10	Hospital data Set	85.7143	75	81.8182	0.8571	6	5.1429

Table 3. Performance Comparison of Proposed ECDC Framework with Conventional Methods (Values are expressed in %)

S. No.	Data Sets	Percentage of Missing Data	Conventional Methods			Proposed ECDC Framework for Data Cleaning (D)	Remarks (Percentage improved (E)=(D)-(B)
			Minimum Method (A)	Average Method (B)	Maximum Method (C)		
1	Voter's Data Sets	0.02	12.35	58.86	39.36	86.72	27.86
2	Ionosphere Data Sets	0.05	12.95	56.91	37.02	85.01	28.1
3	Cylinder Bands Data Sets	0.10	11.96	59.07	36.83	85.35	26.28
4	Mushroom Data Sets	0.15	12.57	59.51	36.65	84.79	25.28
5	Adult Data set	0.20	12.89	57.95	36.37	83.81	25.86
6	Credit-approval Data Sets	0.22	12.91	58.82	39.35	84.85	26.03
7	NLBB Salary Data Sets	0.25	12.89	57.69	36.86	83.38	25.69
8	Hospital Data set	0.30	11.92	58.08	37.82	84.35	26.27

Table 4. Noise data cleaning performance

Sl. No.	Name of the Data sets	% of Noise	Conventional Average Method Performance	ECDC framework Performance
1.	Voter' s Data sets	0.02	58.86	86.72
2.	Ionosphere Data sets	0.05	56.91	85.01
3.	Cylinder Bands Data sets	0.1	59.07	85.35
4.	Mushroom Data sets	0.15	59.51	84.79
5.	Adult Data sets	0.2	57.95	83.81
6.	Credit-approval Data sets	0.22	58.82	84.85
7.	NLBB Salary Data sets	0.25	57.69	83.38
8.	Hospital Data sets	0.3	58.08	84.35

The percentage of noise values were found as 0.02 in Voter's data set, 0.05 in Ionosphere data set, 0.10 in Cylinder band data set, 0.15 in Mushroom data set, 0.20 in Adult data set and 0.22 in Credit-approval data set and 0.25 in NLBB salary dataset and 0.30 in Hospital data set. The proposed ECDC framework detected and eliminated the noise values and replaced the missing values better than the conventional data cleaning methods of Minimum, Average and Maximum.

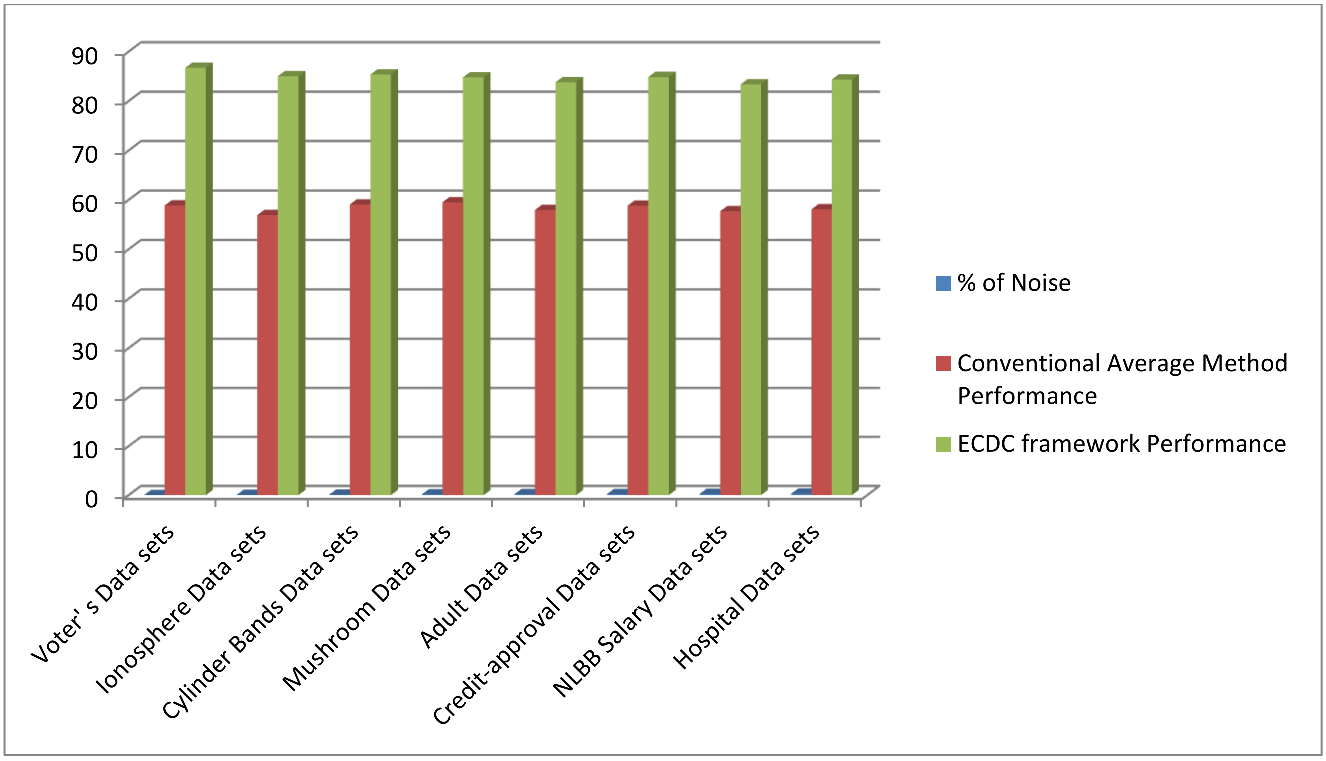

Figure 2. Performance wise Comparison chart of Proposed ECDC Framework with Conventional Average Method

IV. CONCLUSION

This research exercise adopts the implementation of existing data cleaning methods to improve the efficiency of data cleaning methods in data warehouse applications. In this work, the first phase is used to select the data and it would be selected from the multiple data source, then Schema level problem will be rectified using modified XML constraints, and in the second phase, instance level problem will be rectified using ETL based data cleaning with smart tokens. In the third phase, missing values will be replaced by Neural Network then quality data will be loaded into the data warehouse for further process.

In this research 10 datasets were experimented viz., Iris, Voter's data set, Ionosphere data set, Adult data sets, Cylinder band data set, Mushroom data set, Employee data set and NLBB salaries salary data set. The results have improved significantly compared to the conventional data cleaning methods. Hence this research work paves way for researchers to create an ECDC data cleaning tool for all the issues of single and multi-data source based on the proposed ECDC framework.

V. REFERENCES

[1] Elgamal, F, Mosa, NA &Amasha, NA 2014, 'Application of Framework for Data Cleaning to Handle Noisy Data in Data Warehouse', International Journal of Soft Computing and Engineering (IJSCE) vol.3, issue 6, pp.226-231.

[2] Elmagarmid, AK, Ipeirotis, PG &Verykios, VS 2007, 'Duplicate Record Detection', A Survey. IEEE TKDE, vol.19, no.1, pp.1-16.

[3] Erhard Rahm & Hong-Hai Do 2000, 'Data Cleaning: Problems and Current Approaches', IEEE Bulletin of the Technical Committee on Data Engineering, vol.23, no.4, pp.1-10.

[4] EshrefJanuzaj&VisarJanuzaj 2009, 'An Application of Data Mining to Identify Data Quality Problems', In IEEE Proceedings of International Conference on Advanced Engineering Computing and Applications in Sciences, pp.17-22.

[5] Ezeife, CI & Timothy E Ohanekwu 2005, 'Use of Smart Tokens in Cleaning Integrated Warehouse Data', the International Journal of Data Warehousing and Mining (IJDW), vol. 1, no.2, pp. 1-22, Ideas Group Publishers.

[6] Ezeife, CI 2001, 'Selecting and materializing horizontally partitioned warehouse views',

Elsevier Journal of Data and Knowledge Engineering, vol.36, no.2, pp.185-210.

[7] JebamalarTamilselvi, J &Saravanan, V 2008, 'A Unified Framework and Sequential Data Cleaning Approach for a Data warehouse', IJCSNS International Journal of Computer Science and Network Security, vol.8, no.5, pp. 117-121.

[8] JebamalarTamilselvi, J &Saravanan, V 2008, 'Handling Noisy data using Attribute Selection and Smart Tokens', International Conference on Computer -Science and Information Technology, IEEE pp.770-774.

[9] JebamalarTamilselvi, J &Saravanan, V 2010, 'Token-based method of blocking records for large data warehouse', Advances in Information Mining, ISSN: 0975–3265, vol.2, pp.05-10.

[10] Kavitha Kumar, R &Chandrasekaran, RM 2011, 'Attribute correction-data cleaning using Association rule and clustering methods', International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.2, pp.22-32.

[11] Kavitha, PT &Sasipraba, T 2011, 'Performance Evaluation of Algorithms using a Distributed Data Mining Frame Work based on Association Rule Mining', International Journal on Computer Science and Engineering (IJCSE), vol. 3 no. 12 , pp.3845 – 3853

[12] SumonShahriar&SarawatAnam 2009, 'Towards Data Quality and Data Mining using Constraints in XML', International Journal of Database Theory and Application, vol. 2. no. 1, pp.1-8.

[13] Taghi M Khoshgoftaar& Jason Van Hulse2009,' Empirical Case Studies in Attribute Noise Detection', IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 39, no. 4.

[14] TaoxinPeng 2008, 'A Framework for Data Cleaning in Data Warehouses', Napier university, Edinburgh, UK, 1-6. Proc. of the 10th International Conference on Enterprise Information Systems (ICEIS), pp.473-478.

[15] Wei Wei 2001, 'Data mining using Neural Networks for Large Credit Card Record Sets', A MS-Thesis of New Jersey Institute of Technology, New York, pp.1-83.

[16] Wei–Sen Chen & Yen-Kuan Du 2009, 'Using Neural Networks and Data mining Techniques for the financial distress prediction model', Expert System with applications, Elsevier, vol.36, pp.4075-4086.

[17] Xianjun Ni 2008, 'Research of Data Mining Based on Neural Networks', World Academy of Science, Engineering and Technology.

[18] Yashpal Singh &Alok Singh Chauha (2005 – 2009), 'Neural Networks In Data Mining',Journal of Theoretical and Applied Information Technology pp. 37- 42.

Authors Biography:

Agusthiyar. R. He received his Post Graduate MCA degree from Madurai Kamaraj University, Tamil Nadu, India in 2001, and M. Phil degree from Periyar University, Tamil Nadu, India in 2008 respectively. He was submitted his Ph.D thesis in Data mining at Anna University and he started his teaching profession from 2002 to till date and now he is working as Asst. Professor (Sr.G) in Dept. of Computer Applications, SRM University. He has contributed in 5 international Journal papers and number of international and national conferences. His research interests include Data mining, and Artificial Intelligence.

Prof. Dr. K. Narashiman is the Director for AU TVS Center for Quality Management, Anna University Chennai. India. He is a well-known Teacher, Trainer, Consultant and Researcher in the area of Quality Management. He has been recognized by the state of Bremen, Germany for implementing Quality Management System. Trained with scholarship in JAPAN, he is successfully propagating Quality Management to industrial and academic community.