Proceedings of Symposium on Electrical, Mechatronics and Applied Science 2018 (SEMA'18), pp. 145-146, 8 November 2018

FMCleaner: Automatic detect and repair data error using Functional Dependencies and Machine Learning algorithm

Jesmeen M. Z. H. [1,*], J. Hossen[1], K. Tawsif[1], Md. Armanur Rahman[1], Shohel Sayeed[2]

[1)] Faculty of Engineering and Technology, Multimedia University, Melaka, 75450, Malaysia
[2)] Faculty of Information Science & Technology, Multimedia University, Melaka, 75450, Malaysia

[*]Corresponding e-mail: jesmeen.online@gmail.com

Keywords: Data cleaning; Machine Learning; Integrity constraints

ABSTRACT – Data analytics (DA) technology beneficial for analyzing and predicting, and it also helps to make decisions. In the process of DA one of the challenges are detection and repairing dirty data, where failure to do so can result in incorrect analytics and defective decisions. In this paper, an FMCleaner is proposed and implemented to automatically detect (such as Integrity Constraints) and repair data error. The developed system evaluated by using a test dataset with an accuracy of around 95%.

1. INTRODUCTION

Data quality became a major concern for many organizations. As stated by [1] "Data cleaning, as an essential aspect of quality assurance and a determinant of study validity, should not be an exception". Few reasons for producing an error is dataset are missing data, different formats (such as date format), replicated entered data, typos, outlier data and violating business rules.

For the past few years, on emerging new data, industries and academics fields become worried about data cleaning [2–4]. For different format of data required different scripting. However, Machine Learning (ML) have ability to develop an automatic system for cleaning data [5]. Whereas, integrity constraints (ICs) helps to elaborate rules of data cleaning

Contributions. The main contributions in this work are to present current technique for progressive detecting data error and cleaning data automatically using ML as highlighted in figure 1. Moreover, evaluation of the system was executed using four data sets obtained from UCI repository and Malaysia TM Company and it results in better prediction accuracy. Finally, in this paper, result and discussion one of the dataset outcomes discussed.

2. DATA CLEANING TECHNIQUES

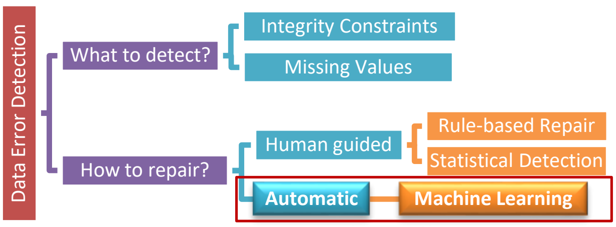

Figure 1 Proposed system data error detection

As shown in figure 1, the system proposed contains two main phases: detection and repairing.

2.1 Error Detection

Detecting IC depending on first-order logic including Functional Dependencies (FDs) to understand the rules of captured data quality. While missing values can also produce issues in analytical process. Designing ICs can obtain manually, but we had proposed a system, which will execute FDs automatically.

2.2 Error Repair

The proposed approaches for repairing dirty data involving ML for predicting missing values, rather than setting missing values in statistic method (such as mean/median). Other methods of cleaning data involved human guidance to confirm the fixes, suggestion to fix or to select the best ML models to process automatic repairing decisions [6].

3. PROBLEM FORMALIZATION

Set of Rules (R) obtained to use for identifying a set of inconsistent values (such as errors due to typos) for each categorical fields.

Pseudocode to discover FD using Pruning Algorithm

```
Inputs:  D is the training set
         Col is columns of the dataset
Output:  Function Dependencies for input dataset
rulesFD={}
C0:={}//Emply list of coloums
C1:= Col//is a copy of listofcolumns
c:=1
C:=[C0,C1]
while Cc!=0
  compute_dependencies(Cc, C1)
  PRUNE(Cc)//helps to Reduce column combinations
  temp:= generate_next_level(Cc)
  Cc+1.append(temp)
  c:=c+1
end
```

Trained ML model used for numerical and non-numerical fields to predict missing values. In this case, we assumed the training set is already clean (Dclean) manually by inspecting the data as far as possible. The system contains list of function F = {f1,…,fn} to repair detected data issues. Considering, DDirty to be set of containing data error and repair with selected function from F.

The implemented algorithm is as follows:

1. Let C_{null} be the missing value columns, Detect C_{null} containing missing values using F_{NAN} (Function to detect is column contains null/NaN/?)
2. Obtaining, D_{clean} (the dataset cleaned previously), and let C_{clean} is list of labels/features/attributes labels for each record containing clean data

3. Train Random Forest (RF) Classifier with selected data set and labels
   Train(Dclean,Cclean)
   if obtained best features
   go to step 6
   else go to next step
4. Extract important features/labels using Gini index obtained from RF. Gini index defined as

$$gini(C) = \sum_{i=1}^{n_l} p_i(1-p_i)$$

   Where n_l is the number of classes in set C (the target variable) and p_i refers ratio of this class i.
5. Repeat step 3
6. Applying the trained classifier to the predict the missing values for set of rows, P=Predict(D_{Dirty})
7. Predicted P data is appended to the D_{clean} data
   Dclean = Dclean ∪ P
8. repeat the process from step 3

Implementation. FMCleaner's script is implemented in Python while data retrieved using "pandas" library and to present outcome "matplotlib" was imported. Development and evaluation executed on Intel CORE i5, 8GB RAM, running Windows 10.

4. Results and Discussion

After implementing FD and few rules, such as is the row contains missing value(s), for detection. For repairing, for categorical and string data ML classification trained model is used to obtain missing value. For numerical data, regression trained model is used to predict closest value possible. For testing the algorithm, rules obtained using a demo data set created as shown in figure 2

	A	B	C	D	E
1	FName	LName	City	Country	PostalCode
2	Jesmeen	Hoque	Melaka	Malaysia	75459
3	Jesmeen	Hoque	Melaka	Malaysia	75450
4	Jesmeen	Hoque	Melaka	Malaysia	75450
5	Tawsif	Hoque	Melka	Malaysia	75450

Figure 2(a) Test data

```
List of all FDs: ....., ['C', 'D'], ['E', 'D']]
Total number of FDs found:  15
```

Figure 2(b) Test data outcome

The outcome shows,

$$\text{rulesFD}[1] = f1: [\text{City}, \text{Country}] \rightarrow PostalCode$$
$$\text{rulesFD}[2] = f2: [\text{PostalCode}, \text{Country}] \rightarrow City$$

Dataset. To evaluate real-world dataset selected from UCI (Diabetics Data [7]) with missing values and errors. One of the results as presented and discussed here.

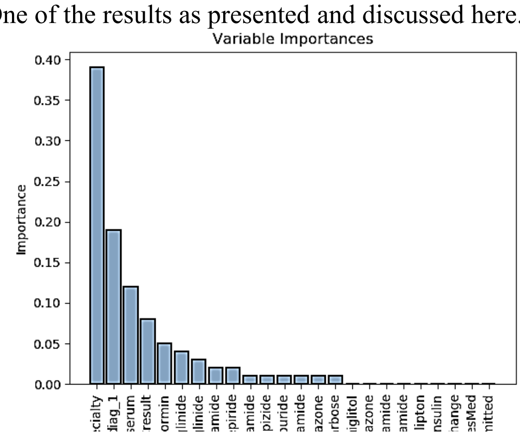

Figure 3. Feature importance for column rosiglitazone.

Missing value for C=rosiglitazone. 101767 rows of data with 25 variables was divided into two sets as following:
Training Features Shape: (76325, 24)
Training Labels Shape: (76325,)
Testing Features Shape: (25442, 24)
Testing Labels Shape: (25442,)
For one of the missing value "rosiglitazone" the feature importance shown in figure 3, 'n' number of features selected which brings 95% importance

After selecting the best features, the training and testing feature and labels fitted again into the system and trained model accuracy plotted and presented in figure 4.

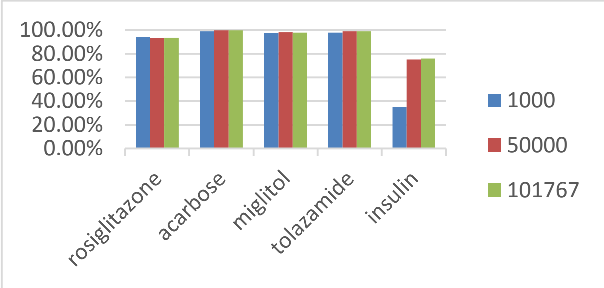

Figure 4. Accuracy of predicting column null value

5. Conclusion

Predictive models are popular in data analytics, but its challenging to manage dirty data. FMCleaner focuses on data detecting and cleaning using constraints and rules. A prototype presented with a new data cleaning system that uses ML technique (i.e. Random Forest) for predicting missing values. Finally, evaluated results on datasets from the UCI repository with real data errors.

REFERENCES

[1] Van Den Broeck, J., Cunningham, S. A., Eeckels, R. and Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities, *PLoS Med.*, *2(10)*, 0966–0970.
[2] Wang, J., Berkeley, U. C., and Tang, N. (2014) Towards Dependable Data Repairing with Fixing Rules, 457–468.
[3] Chu, X., Ilyas, I. F., and Papotti, P. (2013). Holistic Data Cleaning : Putting Violations Into Context, *2013 IEEE 29th International Conference on Data Engineering (ICDE)*
[4] Rekha, M., Marudachalam, N. (2016). Enhanced Data Cleaning Using Ontology Data, *Int J. of Curr Trends in Eng. Research 2(8)*, 140–149.
[5] Jesmeen, M. Z. H., Hossen, J., Sayeed, S., Ho, C. K., Tawsif, K., and Rahman, A. (2018). A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics, *Indones. J. Electr. Eng. Comput. Sci. 10(3)*, 1234–1243.
[6] Yakout, M., Elmagarmid, A. K., Neville, J., Ouzzani, M., and Ilyas, I. F. (2011). Guided Data Repair, *Proc. of the VLDB Endow.*, *4(5)*, 279–289.
[7] UCI (2014). Diabetes 130-US hospitals for years 1999-2008 Data Set, *Clinical and Translational Research, Virginia Commonwealth University*.