Effective Knowledge Discovery Using Data Mining Algorithm

Garima Sharma and Vikas Tripathi

Abstract In present data science world, data is primary for any analytics, analysis, mining, prediction, and description activity. Although many steps are defined and actively used in random for cleaning and preprocessing of dataset previously but there exist some gaps, degrading the overall quality of data as well as knowledge discovery procedure. In this work, some major additional activities are identified in data manipulation paradigm that can enhance decision-making capability of any mining algorithm. With introduction of new methods at collection and cleaning levels, a data preprocessing algorithm is also proposed here. The improvement in overall knowledge discovery process is demonstrated using a real-time dataset.

Keywords Data quality · Data preprocessing · Data analysis · Data cleaning · Knowledge discovery · Data manipulation · R

1 Introduction

Data preprocessing is a broad umbrella which covers ample amount of strategies and techniques that are correlated and interrelated in many ways [1]. For getting zest of any dataset, a major part lies in its cleaning and manipulation of collected data. As nothing can be perfect, so the same problem lies with our data. Before performing any preprocessing activities like aggregation, dimensionality reduction, or feature extraction, quality of data is cynosure [2, 3]. Any analytics results are solely dependent upon the quality of its training dataset.

In real-world data collection, before moving any data into some statistical algorithm like classification, clustering, mining, etc. [4], there are number of pre-requisite steps to be followed for getting rightful, accurate, and trustful results. Extracting new features from given set of attributes is very common nowadays. Though selecting the right attributes from hundreds of given feature set is a matter of expert, keeping the required ones and eliminating the irrelevant or redundant attributes not only helps

G. Sharma (✉) · V. Tripathi
Graphic Era Deemed to be University, Dehradun, Uttarakhand 248002, India
e-mail: garimavrm91@gmail.com

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes
in Networks and Systems 154, https://doi.org/10.1007/978-981-15-8354-4_15

in maintaining the data fitness, reducing the dimensionality of data, as well as helps decision-making algorithms to run faster and more efficiently.

This paper proposed an algorithm for effective knowledge discovery by covering more methods for mitigating data quality issues. Our focus area includes incorporation of new steps in data collection and cleaning, which has direct impact on quality of results in knowledge discovery.

This paper is organized in the sections as in Sect. 2, we discussed about steps to be performed at the time of data collection, so that an advance refinement of collected data could be done at the first level itself along with data analysis and cleaning activities, and we have covered few more checks and treatments to be perform in data preprocessing. The proposed algorithm and detailed implementation strategies are explained in Sect. 3, and Sect. 4 showcases the impact of given algorithm in effective and more truthful knowledge discovery. The concluding remarks and future work is given in Sect. 5 showcasing how one can achieve staggering and more reliable results by stepwise implementing the given algorithm.

2 Related Work

2.1 Data Collection

A classical definition of data collection is gathering of information in a systematic fashion. This statement has evolved a lot with time. At present, data collection tools and techniques are way more than just fetching and loading of data. Complete ETL process—extract, transform, load is expected from a collection tool in modern systems. Keeping this into view, we analyzed few datasets and found some abnormalities other than the implemented ones [5]. As there could be more than one source while fetching data into a particular system, a problem of inconsistency exist in column names, i.e., for a similar attribute, there could exist different name from different sources. Second irregularity we found was distinct formats for a single attribute. Most widely seen example for this case is timestamp. Third anomaly proved primer in erroneous result in knowledge discovery phase was incorrect column type, i.e., the attribute data type was not matching with its values. These anomalies are required to be essentially removed at the time of collection and implicitly before analysis.

2.2 Data Analysis and Cleaning

Data analysis is a process of organization of data in drawing helpful conclusions. This phase acts as base for various data cleaning [4] activities. Identification and removal of inconsistent and imprecise values present in any crude dataset is main aim of any data cleaning method. Noise and outlier detection algorithms like clustering or unsupervised machine learning algorithm work efficiently in searching and removing the abnormal records [6]. Here, abnormal refers to the outliers or oddly present dataset showing serious deviation from other data items present with in the dataset. As unsupervised algorithms has no labels, and therefore, no boundaries exist for framing the data items, thus helpful in finding anomalies. This is classically performed in all the analysis work [7]. Here, we find an improvement in this legacy analysis and cleaning system. As we know, there could exists hundreds, thousands of features in a single dataset [8]; therefore, there may exist a possibility of interconnection and interrelation between them. These relationships can be used in treating missing values as well as NULL values present in the dataset. These relationship values so obtained perform a crucial role in data manipulation. Various machine learning algorithms like apriori algorithm and KNN can be used to predict the missing values using these relationships. Redundancy in records are required to be removed after all data manipulation.

2.3 Data Preprocessing

To get data finally ready for discovering knowledge, it must be passed through data preprocessing phase. This generally includes integration of various attributes and creation of a new attributes by aggregation or segregation of attributes, selecting required and primary features while dropping irrelevant ones [5]. Normalization can be done in the end of preprocessing unit if in case scaling is required. This can be performed in two ways—min-max normalization and z score normalization [9].

3 Proposed Algorithm

Following is the associated algorithm to be executed stepwise for getting maximum data quality and knowledge discovery (Fig. 1):

1. Ingest dataset (K_i) whereas 'i' is the number of sources, 'n' is the number of columns, and 'r' is the number of rows.
2. Check and rectify column name mismatching between similar attributes of different sources and make one unit by combining all the datasets (K).
3. Now, further check number of columns (n) and their formats, i.e., whether all the values are present in a single format or not. If no, correct the same and proceed to step 04
4. Detect the datatype of each column and again check whether they are complementing with respective column values. If no, correct the same and proceed to step 05
5. Start data analysis phase by detecting and removing noise and outliers present in the dataset.

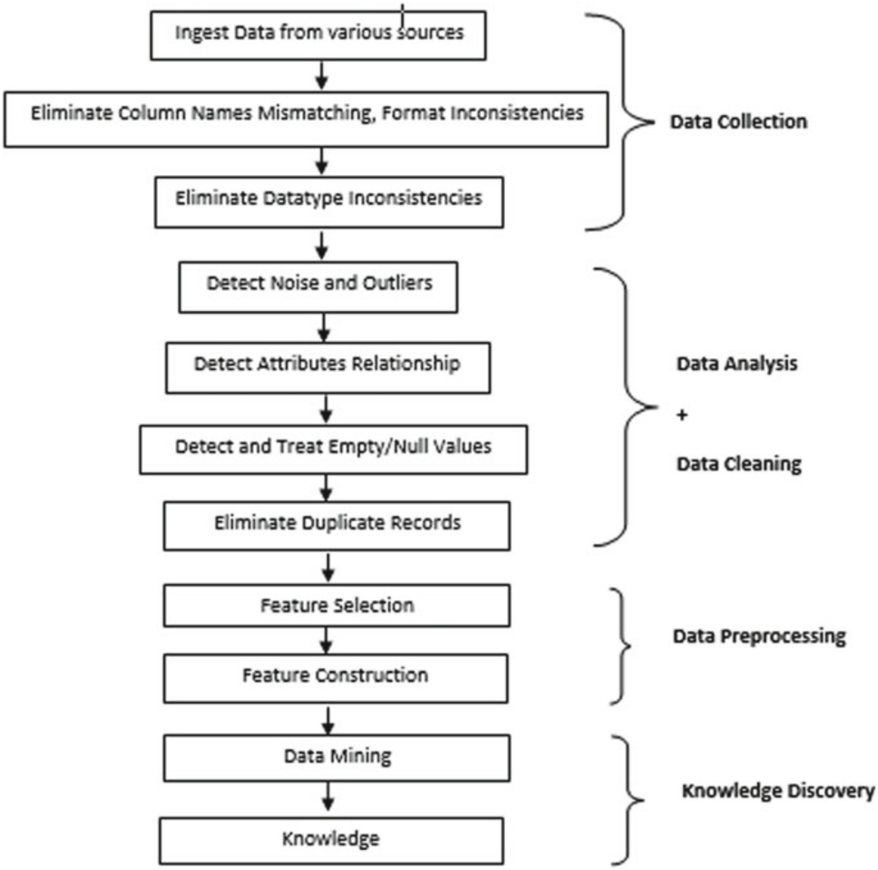

Ingest Data from various sources
Eliminate Column Names Mismatching, Format Inconsistencies
Eliminate Datatype Inconsistencies
Data Collection
Detect Noise and Outliers
Detect Attributes Relationship
Detect and Treat Empty/Null Values
Eliminate Duplicate Records
Data Analysis
+
Data Cleaning
Feature Selection
Feature Construction
Data Preprocessing
Data Mining
Knowledge
Knowledge Discovery

Fig. 1 Proposed data mining algorithm

6. Since there is only one dataset now, detect the relationships between the different attributes. This is useful in data manipulation, i.e., for treating missing and NULL values. Rectify and correct such values using this step.
7. Check duplicate values ‘dR’ present in the dataset (K), if yes, then go to step 08, else go to step 09.
8. Remove duplicates and check the unique number of rows ‘UR’

$$r = \mathrm{dR} + \mathrm{UR}$$

9. Proceed further with other data preprocessing activities like feature selection and construction of new attributes from given attributes by following aggregation, segregation, etc.
10. Finally, knowledge discovery procedure can be begin based upon the use cases.

4 Knowledge Discovery

Knowledge discovery is solely dependent upon the quality of data passing into discovery systems [10]. This involves application of various data preprocessing methods aimed at facilitation of data mining algorithms. Many times even required post-processing for refining and improvement of knowledge [11]. For validation of stated algorithm, we have taken a dataset from NYC open data Web site [12]. The data contains information of dog owners living in New York City. All the residents of New York City are required to license their dogs right after their adoption as per the given law. In our dataset, each record represents a unique dog license issued date and expiry date. Each tuple stands as a unique license period for a dog over a year-long time period. This dataset has 15 columns and 51,861 rows saved in a csv format. After analyzing the dataset attributes, other than problems like null value and missing value, there were major data quality issues, refer Table 1.

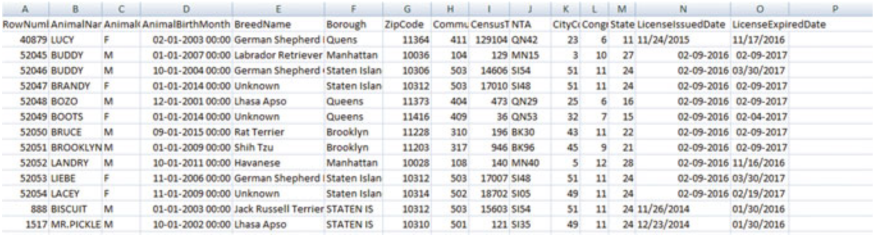
To understand, the dirty data here is screenshot of our sample grubby dataset, and its inconsistencies are discussed in Table 1 quality issues column (Fig. 2).

The main aim of this work is to showcase the importance of new methods explored and implemented in data collection and cleaning level. The results shows how the new algorithm is impacting the overall knowledge discovery procedure of given dataset. We have used an open-source statistical language, R, and analyzed the results using exploratory data analysis technique [15].

In Fig. 3, due to presence of imprecise boroughs, the distribution checkup between the NYC states was incorrect, but after finding the relationship between zipcode and

Table 1 Data quality issues and suggested solution

Column(s) name	Messy area	Suggested solution	Quality issue
Animal gender	03 Types—M, F, “ ”	Drop the missing value row	Data inconsistency
Animal birth name	String datatype	Timestamp datatype	Data type inconsistency
Borough	New York City has three different abbreviations Queens has two different abbreviations Staten Islands has two different abbreviations	Single name to one area type	Misspellings
ZipCode	Correct zipcode maps to number of “unknown” boroughs Same zipcode maps to different boroughs	Zipcode–borough mapping correction using correct data [13, 14]	Find relationship between zipcode and borough
License issued date	String datatype	Date datatype	Incorrect datatype
License expired date	String datatype	Date datatype	Incorrect datatype



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
RowNuml	AnimalNar	Animal	AnimalBirthMonth	BreedName	Borough	ZipCode	Commu	CensusT	NTA	CityC	Congi	State	LicenseIssuedDate	LicenseExpiredDate	
40879	LUCY	F	02-01-2003 00:00	German Shepherd	Quens	11364	411	129104	QN42	23	6	11	11/24/2015	11/17/2016	
52045	BUDDY	M	01-01-2007 00:00	Labrador Retriever	Manhattan	10036	104	129	MN15	3	10	27	02-09-2016	02-09-2017	
52046	BUDDY	M	10-01-2004 00:00	German Shepherd	Staten Islan	10306	503	14606	SI54	51	11	24	02-09-2016	03/30/2017	
52047	BRANDY	F	01-01-2014 00:00	Unknown	Staten Islan	10312	503	17010	SI48	51	11	24	02-09-2016	02-09-2017	
52048	BOZO	M	12-01-2001 00:00	Lhasa Apso	Queens	11373	404	473	QN29	25	6	16	02-09-2016	02-09-2017	
52049	BOOTS	F	01-01-2014 00:00	Unknown	Queens	11416	409	36	QN53	32	7	15	02-09-2016	02-04-2017	
52050	BRUCE	M	09-01-2015 00:00	Rat Terrier	Brooklyn	11228	310	196	BK30	43	11	22	02-09-2016	02-09-2017	
52051	BROOKLYN	M	01-01-2009 00:00	Shih Tzu	Brooklyn	11203	317	946	BK96	45	9	21	02-09-2016	02-09-2017	
52052	LANDRY	M	10-01-2011 00:00	Havanese	Manhattan	10028	108	140	MN40	5	12	28	02-09-2016	11/16/2016	
52053	LIEBE	F	11-01-2006 00:00	German Shepherd	Staten Islan	10312	503	17007	SI48	51	11	24	02-09-2016	03/30/2017	
52054	LACEY	F	11-01-2009 00:00	Unknown	Staten Islan	10314	502	18702	SI05	49	11	24	02-09-2016	02/19/2017	
888	BISCUIT	M	01-01-2003 00:00	Jack Russell Terrier	STATEN IS	10312	503	15603	SI54	51	11	24	11/26/2014	01/30/2016	
1517	MR.PICKLE	M	10-01-2002 00:00	Lhasa Apso	STATEN IS	10310	501	121	SI35	49	11	24	12/23/2014	01/30/2016	

Fig. 2 Screenshot of grubby dataset opened in excel

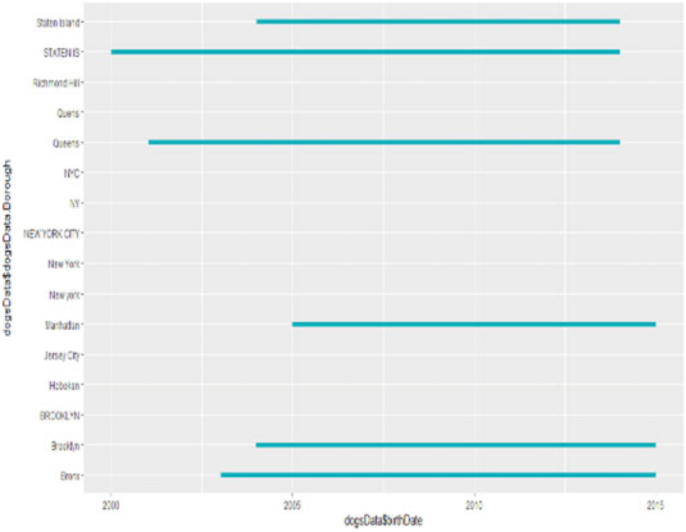


Fig. 3 Number of dogs/town without cleaning

borough (city name), we were able to correct the missing and incorrect boroughs which eventually impact our distribution of dogs in a particular town, refer Fig. 4.

Also, it was not possible to plot a comparison report between birth dates and license issue dates as both the columns were present in string format with different date format types. After correcting the datatypes at data collection time, we were able to set a contrast chart between the two columns, and the results can be seen in Fig. 5.

As number of boroughs were present, it was not feasible to plot a gender distribution chart for all the boroughs individually. After data preprocessing, we were able to set a contrast chart between the boroughs in Fig. 6.

Using Fig. 6, we can even discover information about number of males or females present in a particular borough, we can calculate the ratio between the males and females in a particular town, etc.

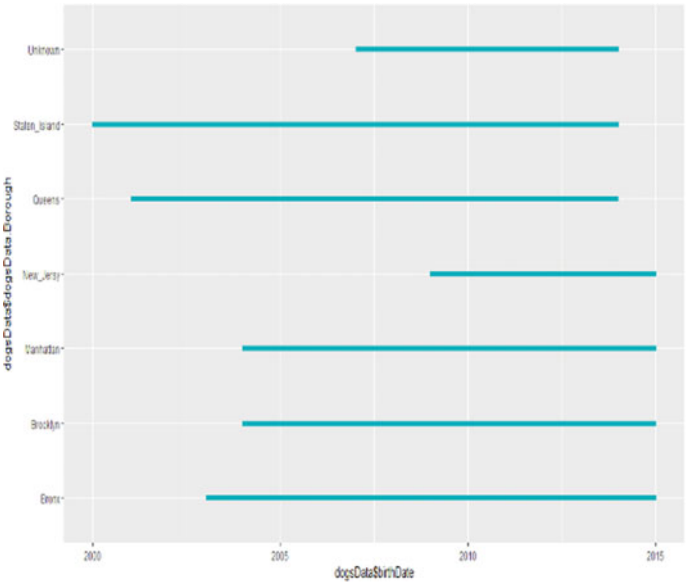


Fig. 4 Number of dogs/town after cleaning

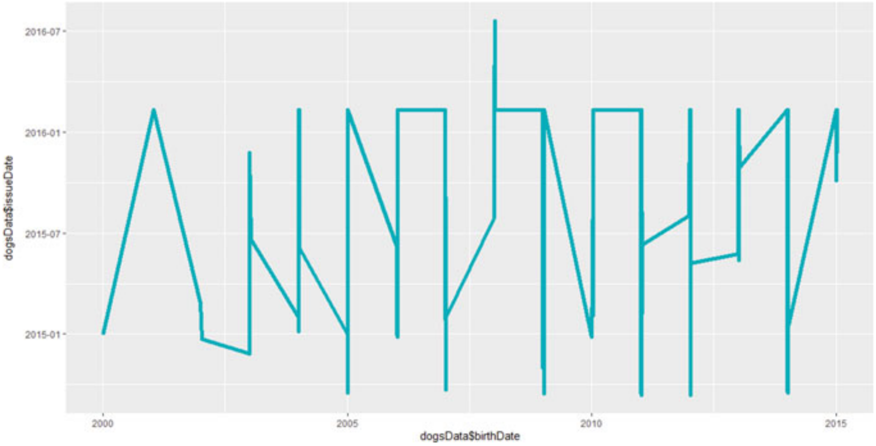


Fig. 5 Dogs birth date versus license issued dates

Lastly, we wished to obtain most favorite breed of New York City, and then again it was not possible with raw data due to presence of enormous number of null values. After treating the null values by setting relationship between the column and dropping all the unknown breeds. After cleaning and preprocessing the transformed dataset, we obtain following word cloud based on the number of counts. Higher the count, more centered the position of value.

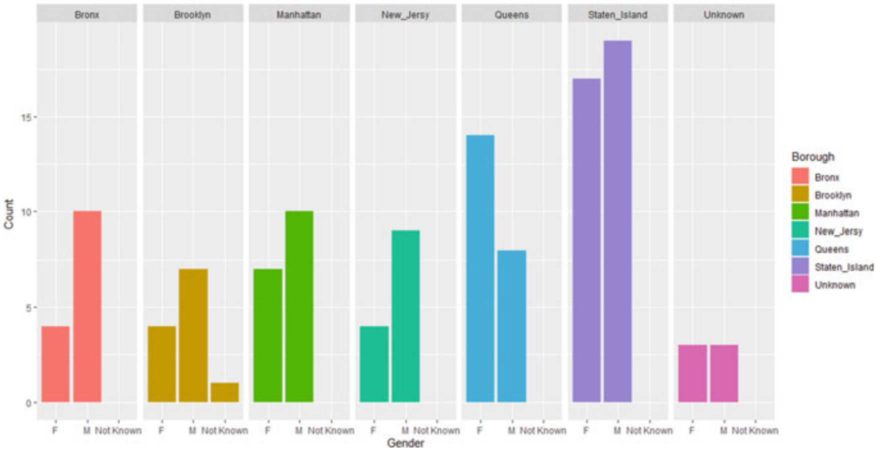


Fig. 6 Gender distribution with respect to boroughs after cleaning

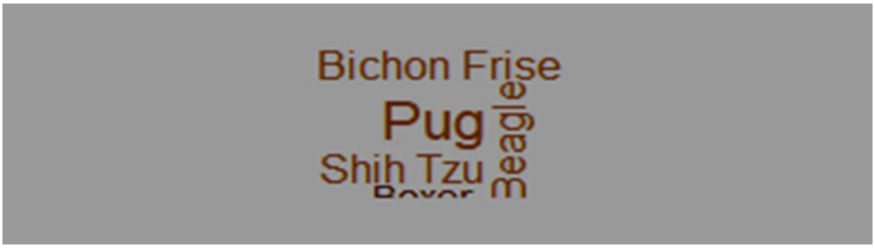


Fig. 7 Favorite dog breeds after data cleaning

Therefore, pug is the most favorite dog in whole NYC region of The United States (Fig. 7).

5 Conclusion and Future Work

Data mining is the process of discovering useful information in a large data repository [11]. This single phase requires number of pre-requisite activities to be followed in a sequence. In our work, we have covered all data mining activity level with inclusion of new activities to improve the knowledge discovery procedure. We carefully analyze the deep insights for data collection and its preprocessing units and suggested an algorithm for effective mining of textual dataset. This algorithm can also be useful in enhancing the over all data quality of any analytics system. The results section demonstrates each step of the proposed algorithm and shows the fruitful impact on dataset quality and knowledge discovery. This has capability of extension, if any new abnormality is found in future. More explorations can be done on cleaning requirements of textual datasets using fusion of machine learning algorithms. It would be nice if a single sequence of this data mining algorithm gives best performance for each type of dataset.

References

1. Grzymala-Busse, J. W., Grzymala-Busse, W. J.: Handling missing attribute value. In O. Maimon & L. Rokach (Ed.), *Data mining and knowledge discovery handbook* (pp. 573–589). Springer.
2. Logan, C., Parás, P., Robbins, M., & Zechmeister, E. J. (2020). Improving data quality in face-to-face survey research. *PS: Political Science & Politics, 53*(1), 46–50.
3. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal, 14.*
4. García, S., Ramírez-Gallego, S., Luengo, J., et al. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics, 1,* 9.
5. García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Intelligent Systems Reference Library.
6. García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Science Direct Knowledge-Based Systems, 98,* 1–29.
7. Oni, S., Chen, Z., Hoban, S., & Jademi, O. (2019). A comparative study of data cleaning tools. *International Journal of Data Warehousing and Mining, 15*(4), 48–65.
8. Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information & Management, 57*(1), 103146.
9. Alasadi, S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences, 12,* 4102–4107.
10. Liu, Q., Xiao, F., Zhao, Z. (2020). *Grouting knowledge discovery based on data mining. Tunnelling and Underground Space Technology, 95*. ISSN 0886-7798.
11. Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (2nd ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
12. NYC Open Dataset. (2018). https://data.cityofnewyork.us/browse?q=NYC+Dogs+Licenses.
13. Postal Codes of New York City. (2019). https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm.
14. Postal Codes of The United States of America. (2019). https://www.postalpinzipcodes.com/Postcode-USA-United-States-ZIP-Code-8527-Postal-Code.
15. Cordón, I., Luengo, J., García, S., Herrera, F., & Charte, F. (2019). Smartdata: Data preprocessing to achieve smart data in R. *Neurocomputing, 360,* 1–13.