International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Data Cleaning Techniques for Large Data Sets

Yogita Bansal, Ankita Chopra

Abstract*: *In today's emerging era of data science where data plays a huge role for accurate decision making process it is very important to work on cleaned and irredundant data. As data is gathered from multiple sources it might contain anomalies, missing values etc. which needs to be removed this process is called data pre-processing. In this paper we perform data pre-processing on news popularity data set where extraction , transform and loading (ETL) is done .The outcome of the process is cleaned and refined news data set which can be used to do further analysis for knowledge discovery on popularity of news . Refined data give accurate predictions and can be better utilized in decision making process.

Keywords: Data Mining, Data Pre Processing, Extraction, Transform, load, Knowledge discovery.

I. INTRODUCTION

Data is very important for accurate business decision making process. In today's times data is gathered from multiple sources and they are scattered and in different formats. To make the data relevant for knowledge discovery it is very much needed to pre-process the data by cleaning & normalizing it. Data cleaning is a step in KDD process .In this paper, News popularity data set is taken in which there are 61 attributes which provide information on number of words in content, news on topics such as entertainment, business, social media, best keyword, average keyword & worst keyword and so on. As the data set is huge with many attributes it needs to be pre-processed before it is used for analysis. In order to accomplish this python libraries are used appropriately wherever required.

II. METHODOLOGY USED

In this paper we are using python libraries for data cleaning. Data cleaning is a step in pre-processing where the data is prepared for analysis and decision making.

A. Data Mining

The procedure of deriving knowledge from given information is called Data Mining. It is observed as a step in KDD process.

According to figure 1,KDD process is shown where the data is first collected from multiple sources in heterogenous formats and sent for preprocessing.

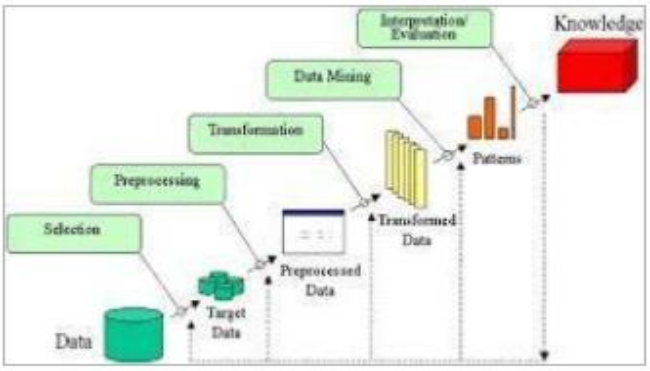

Fig 1: KDD Process

In preprocessing step the entire data is cleaned and transformed into normalized values, as shown in figure 2. In the data set we considered data cleaning is done using python libraries to remove the missing values and keep it clean for further analysis. Once the data is cleaned it is further transformed using PCA(principal component analysis) and the preprocessed data is now ready for application of algorithms.

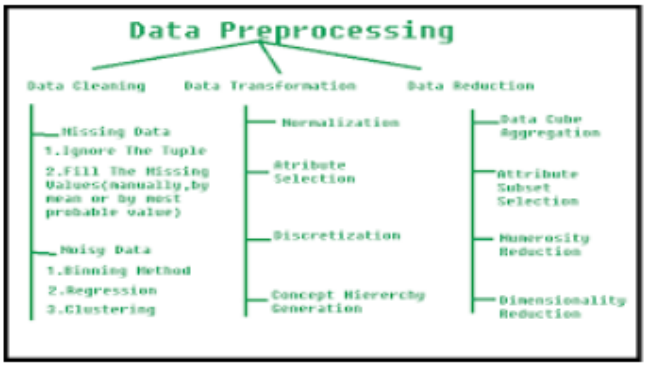

Fig 2: Pre-processing process

In the figure 2, we try to demonstrate essential data pre-processing steps:

Data Cleaning, Data Transformation &Data Reduction: In data cleaning is important because as the data may contain missing values, noisy data and not cleaning them would yield incorrect patterns.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Yogita Bansal, MCA department, Jagan Institute of Management studies, Delhi, India. E-mail: yogita.sharma@jimsindia.org
Ankita Chopra*, MCA department, Jagan Institute of Management studies, Delhi, India. E-mail: ankita.chopra@jimsindia.org

There are various methods for data cleaning they are binning method, regression etc. Once the data is cleaned it needs to be transformed by using normalization & discretization as the values need to fall in uniform range for the results to be appropriate.

After transformation we need to reduce the data dimensionality by numerosity reduction and dimensionality reduction. Thus after applying all the above methods the data is now ready for application of algorithms to yield the results.

Though, it looks like an easy step but preprocessing involves lot of activities. In this paper we try to implement ETL strategies on news data set by using ANACONDA –JUPYTER as tool with python programming.

III. EXPERIMENTS SETUP

The dataset is formerly attained and pre-processed by K. Fernandes et al.[13]. The dataset has 61 attributes (as numerical values) describing various characteristics of each news article, from a total of 39,644 articles. Figure 3 provides a full representation of this feature set. Using this setup, we may hypothesis a lot regarding the data set. Typically, we may model this scenario as a regression problem:

Aim: Given the set of attributes for a typical news article, predict the "number of shares" that the editorial will get when published.

0. **url:** URL of the article (non-predictive)
1. **timedelta:** Days between the article publication and the dataset acquisition (non-predictive)
2. **n_tokens_title:** Number of words in the title
3. **n_tokens_content:** Number of words in the content
4. **n_unique_tokens:** Rate of unique words in the content
5. **n_non_stop_words:** Rate of non-stop words in the content
6. **n_non_stop_unique_tokens:** Rate of unique non-stop words in the content
7. **num_hrefs:** Number of links
8. **num_self_hrefs:** Number of links to other articles published by Mashable
9. **num_imgs:** Number of images
10. **num_videos:** Number of videos
11. **average_token_length:** Average length of the words in the content
12. **num_keywords:** Number of keywords in the metadata
13. **data_channel_is_lifestyle:** Is data channel 'Lifestyle'?
14. **data_channel_is_entertainment:** Is data channel 'Entertainment'?
15. **data_channel_is_bus:** Is data channel 'Business'?
16. **data_channel_is_socmed:** Is data channel 'Social Media'?
17. **data_channel_is_tech:** Is data channel 'Tech'?
18. **data_channel_is_world:** Is data channel 'World'?
19. **kw_min_min:** Worst keyword (min. shares)
20. **kw_max_min:** Worst keyword (max. shares)
21. **kw_avg_min:** Worst keyword (avg. shares)
22. **kw_min_max:** Best keyword (min. shares)
23. **kw_max_max:** Best keyword (max. shares)
24. **kw_avg_max:** Best keyword (avg. shares)
25. **kw_min_avg:** Avg. keyword (min. shares)
26. **kw_max_avg:** Avg. keyword (max. shares)
27. **kw_avg_avg:** Avg. keyword (avg. shares)
28. **self_reference_min_shares:** Min. shares of referenced articles in Mashable
29. **self_reference_max_shares:** Max. shares of referenced articles in Mashable
30. **self_reference_avg_sharess:** Avg. shares of referenced articles in Mashable
31. **weekday_is_monday:** Was the article published on a Monday?
32. **weekday_is_tuesday:** Was the article published on a Tuesday?
33. **weekday_is_wednesday:** Was the article published on a Wednesday?
34. **weekday_is_thursday:** Was the article published on a Thursday?
35. **weekday_is_friday:** Was the article published on a Friday?
36. **weekday_is_saturday:** Was the article published on a Saturday?
37. **weekday_is_sunday:** Was the article published on a Sunday?
38. **is_weekend:** Was the article published on the weekend?
39. **LDA_00:** Closeness to LDA topic 0
40. **LDA_01:** Closeness to LDA topic 1
41. **LDA_02:** Closeness to LDA topic 2
42. **LDA_03:** Closeness to LDA topic 3
43. **LDA_04:** Closeness to LDA topic 4
44. **global_subjectivity:** Text subjectivity
45. **global_sentiment_polarity:** Text sentiment polarity
46. **global_rate_positive_words:** Rate of positive words in the content
47. **global_rate_negative_words:** Rate of negative words in the content
48. **rate_positive_words:** Rate of positive words among non-neutral tokens
49. **rate_negative_words:** Rate of negative words among non-neutral tokens
50. **avg_positive_polarity:** Avg. polarity of positive words
51. **min_positive_polarity:** Min. polarity of positive words
52. **max_positive_polarity:** Max. polarity of positive words
53. **avg_negative_polarity:** Avg. polarity of negative words
54. **min_negative_polarity:** Min. polarity of negative words
55. **max_negative_polarity:** Max polarity of negative words
56. **title_subjectivity:** Title subjectivity
57. **title_sentiment_polarity:** Title polarity
58. **abs_title_subjectivity:** Absolute subjectivity level
59. **abs_title_sentiment_polarity:** Absolute polarity level
60. **shares:** Number of shares (target)

Figure 3: Feature Set

Also, we may define the problem as a classification problem as following.

Aim: Given the set of attributes for a typical news article predict whether editorial will be widespread or not (Yes or No).

Our aim is to perform setup using data pre-processing steps to make the data ready for one of the above-mentioned models.

IV. RESULTS AND DISCUSSION

For performing any data pre-processing, The initial step is to explore the data to find the scales, distribution or any outliers/missing values and other trends. For example, we may make an assumption about people are not interested in reading longer news articles and hence would not share longer articles. The graph in figure 4 shows a scatter plot between number of words in the content of the editorial and number of shares created for the same.

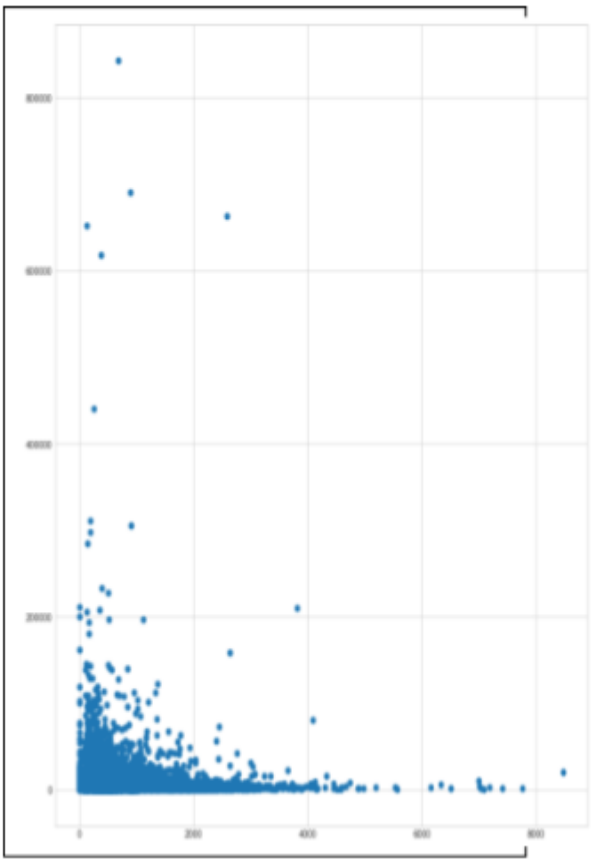

Figure 4: correlation between number of words and shares

This graph supports our assumption of the hypothesis and indicates a negative relationship between the number of words in the content and the number of shares. Another observation is drawn using box plots, which provides information regarding the number of tokens in the title of the editorial grouped by the number of shares, as shown in figure 5, to observe the explained variance in the data.

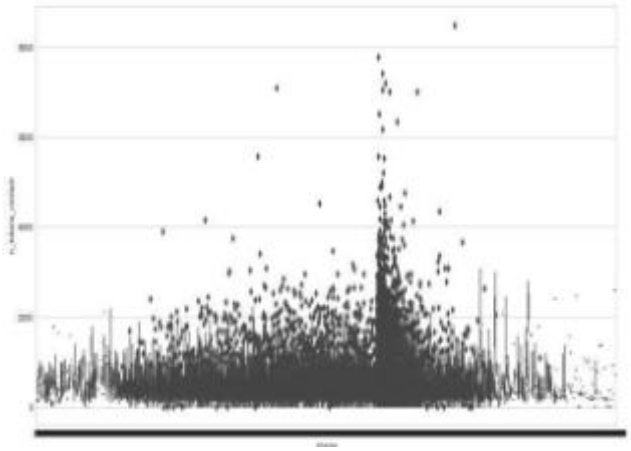

Figure 5: Boxplot to see variance in the data

The analysis shows that the number of tokens in the title is a determining factor about whether the reader would like to read more or not.

It can also be seen from the analysis that the data contains more editorials that are published on weekdays as compared to weekends. This can give two main observations:

1. Data collection for weekends was not enough as compared to data collection for weekdays.
2. A tend to produce less editorials on weekends as compared to weekdays.

Another observation done is to find out the popularity of article, we have divided data on the basis of median no of shares in the dataset. The median comes out to be 1400, so we can make editorial with more than 1400 shares as popular while less than 1400 shares is marked as unpopular.

Figure 6 shows the division of dataset on basis of median no of shares. The data looks balanced and hence can be said to be a good measure for predicting popularity of editorial.

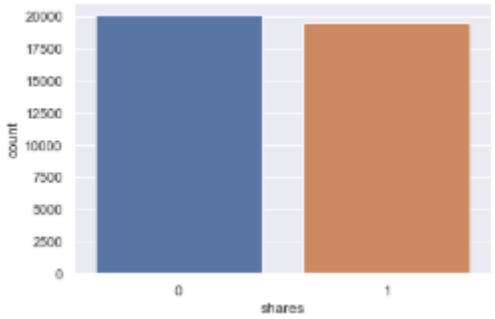

Figure 6: Balanced dataset

The two steps performed to make data ready for application of machine learning models is as follows:

Data Preparing: The attributes namely "URL" and "timedelta" have been discarded since they are meta-data and cannot be treated as features.

Scaling: As different features may hhave different ranges, some features may affect the prediction more than others. Therefore, standardization and normalization of the data is required. Out of many variations of such standardization, we have used Min Max Scaling. This technique scales and translates each feature independently such that all values fall in a given range on the training set, i.e. between zero and one. The dataset is shown in figure 7 as a comparison between before and after the scaling process.

Feature Engineering: This is the most important part of any machine learning model, to get only the attributes that contributes fully in any prediction or regression problem. In our case out of 57 features, a small subset has to be chosen to get the appropriate results. Using Feature importance and dimensionality reduction techniques we found out 40 features were important with importance value lying between [0.0, 1.0]. We removed the features with smaller importance to reduce noise in the data to a great margin. This has been accomplished using Singular value decomposition. We found out that adding 10 components to the list of attributes reduced the noise from data to a greater level.

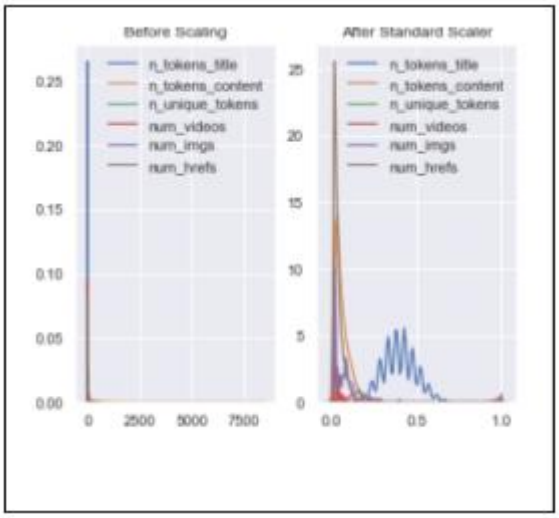

Figure 7: Normalizing data

V. CONCLUSION AND FUTURE SCOPE

Our objective was to pre-process data so that it becomes workable and machine learning models can be applied on it. In this article we tried to elaborate on how one large dataset after preprocessing can be used appropriately to draw insights. In future we will try to apply various techniques for preprocessing to draw comparisons between various strategies available. Also we will try to apply various machine learning models to find out how useful and important are the preprocessing techniques for a better prediction system. Using broad set of techniques we may find out how to use them correctly and analyze which one suits in a particular situation.

REFERENCES

1. N. Katsaras, P. Wolfson, J. Kinsey and B. Senauer, 'Data Mining A segmentation analysis of U.S. grocery shoppers', The Retail Food Industry Center University of Minnesota, March 2001.
2. K. Kumar, B. K. Chauhan, J.P. Pandey, A. K. Tomer, 'Data mining and knowledge discovery from research problems', International Journal of Advanced Technology & Engineering Research (IJATER), 1st International Conference on Research in Science, Engineering & Management (IOCRSEM 2014), Pg. 61-66.
3. B. M. Ramage, 'Data Mining Techniques and Applications', Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305, Pg.301-305.
4. K. M. Raval, 'Data Mining Techniques', International Journal of Advanced Research in Computer Science and Software Engineering', Volume 2, Issue 10, October 2012.

5. L.Rokach and O. Maimon, 'Clustering Methods', Data Mining and Knowledge Discovery Handbook',Pg. 321-353.
6. G.S. Linoff, M. J. A. Berry, 'Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd Edition'.
7. L.Rokach and O. Maimon, 'Data mining for improving the quality of manufacturing: a feature set decomposition approach'.
8. J.Vasilev, 'Data mining of transactional data for sales of dairy products', Theoretical and Applied Economics, Volume XXI (2014), No. 12(601), pp. 3-12.
9. Zainal Fikri Zamzuria, Mazani Manaf, Adnan Ahmad, Yuzaimi Yunus, "Computer Security Threats Towards the ELearning System Assets", Communications in Computers and Information Science, publisher Springer pp: 335-345, Vol- 180 CCIS, ISSN (Print): 18650929, June 2011.
10. Nikhilesh Barik, Dr. Sunil Karforma, "Risks and Remedies in E-learning System," International Journal of Network Security & Its Applications (IJNSA), vol. 4, No. 1, pp. 51-59, Jan 2012, DOI: 10.5121/ijnsa.2012.4105. Available online at: http://arxiv.org/ftp/arxiv/papers/1205/1205.2711.pdf
11. The STRIDE Threat Model: by https://msdn.microsoft.com/enus/library/ee823878(v=cs.20).aspx
12. Ahmad Tasnim Siddiqui , Dr. Mehedi Masud," An E-learning System for Quality Education ",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 ISSN (Online): 1694-0814
13. Fernandes K., Vinagre P., Cortez P. (2015) A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In: Pereira F., Machado P., Costa E., Cardoso A. (eds) Progress in Artificial Intelligence. EPIA 2015. Lecture Notes in Computer Science, vol 9273. Springer, Cham

AUTHORS PROFILE

Ms. Yogita Bansal is working as an Assistant professor in Information Technology department at Jagan Institute of management studies (JIMS).She has pursued her master in Information Technology from IIT Delhi and also has 9 years of versatile experience in industry and academics. Her research area includes Natural Language processing,

Ankita Chopra is currently working as Assistant Professor (IT) with JIMS, Rohini, Delhi. She has a rich experience of 9 years. Her research interest lies in the field of Predictive Data Analytics. she completed her bachelor's and masters in Technology in Computer Science from Jawaharlal Nehru University (JNTU-H) Hyd. She has published several papers in National and International conferences and also published papers in International Journals on Educational Data Mining Techniques, Social Media Analytics, Quantitative Analysis of Dairy Product packaging using Data Mining Techniques.