Technical Seminar

On

Big Data Pre-Processing

Guide Name:
Prof. Sajitha N
Assistant Professor

By:
Ramya shree B
1BG16CS081

Contents

Introduction

- Big data refers to the large, diverse sets of information that grow at every-increasing rates.
- All the sectors like energy,banking,retail,hardware,networking etc all generate huge amount of heterogeneous data.
- Big data helps in acquiring,processing and analyzing large amounts of heterogeneous data to derive valuable results.
- To improve the quantity of the big data we need to pre-process the raw data as it can not be usable as it is.

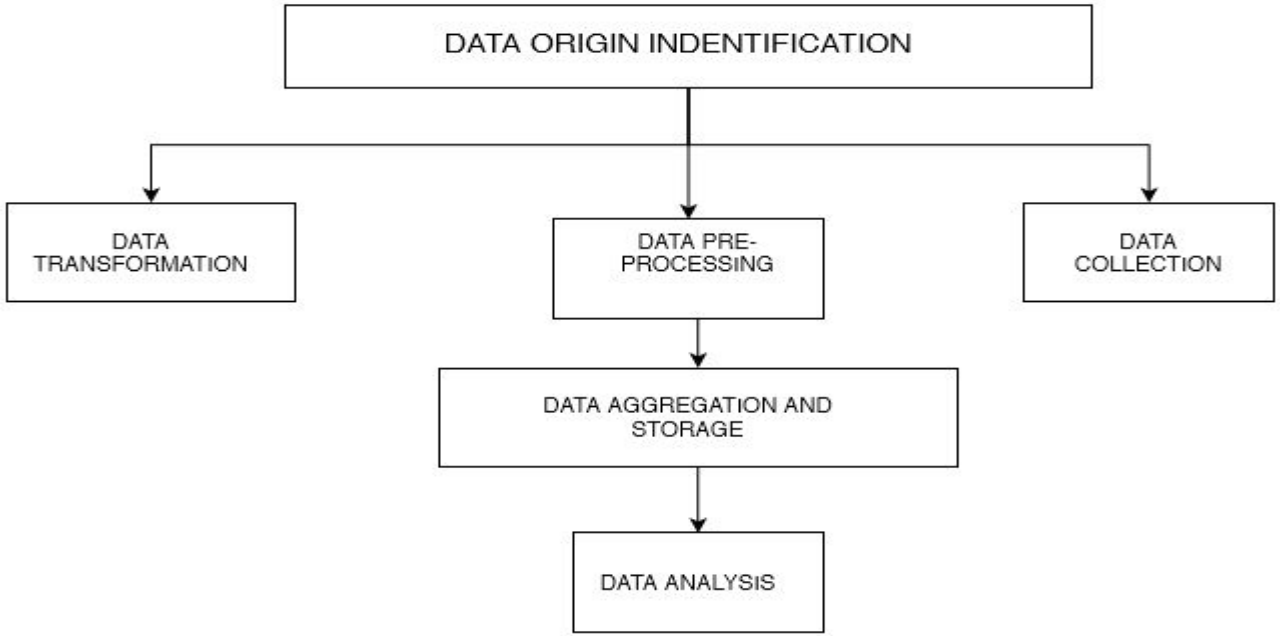
Problem Statement

The raw data available from different sources which if mined, processed and analyzed accurately can improve the quality of the output.

Literature survey

1.Ashish Juneja, Nripendra Narayan Das,Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing.

- Big Data has become as important part of all industries and business sectors today.
- Big data is a large volume of data which comes with complexities commonly known as 5v’s i,e.volume,velocity,variety,veracity,value.
- The Big Data system consist of four phases those are Data origin identification,data cleaning,data aggregation and storage,data analysis.


DATA ORIGIN INDENTIFICATION
DATA TRANSFORMATION
DATA PRE-PROCESSING
DATA COLLECTION
DATA AGGREGATION AND STORAGE
DATA ANALYSIS

The key data dimensions

- Completeness
- Timeliness
- Uniqueness
- Integrity
- Consistency
- Accuracy

Factors of Data Quantity

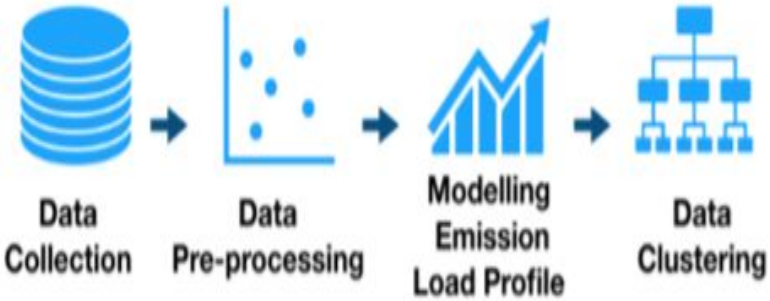
- Data Quality Dimensions.
- Data profiling.
- Data Quality Framework.
- Big Data Pre-Processing.

2.Preeti Nair,Indu Kashyap,Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for kNN Classifier,2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing.

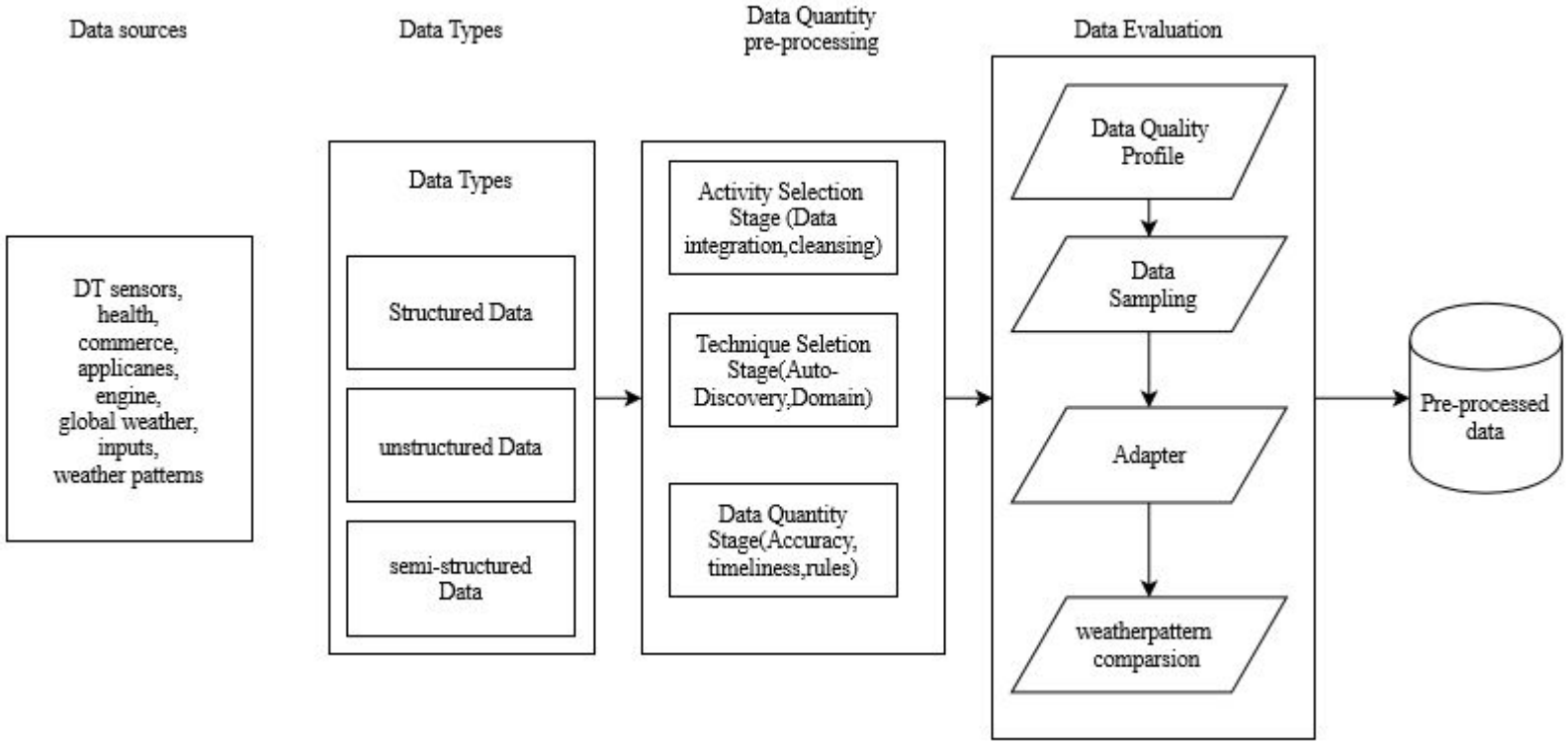
- In data mining, classification means a process which categorizes a collection of data into different groups.
- Hybrid pre-processing technique is a combination of resampling and Interquartile Range (IQR).
- In IQR data is divided into four groups, i.e. the 25th, 50th and 75th values.It is calculated as the variation between the 75th and the 25th percentiles of the entire data under analysis.

3.Angreine Kewo,Pinrolinvic Manembu,Per Sieverts Nielsen,Data Pre-processing Techniques in the Regional Emission's Load Profiles Case,2019 6th International Conference on Control, Decision and Information Technologies.

- Energy consumption is used as a tool for increasing energy efficiency and developing emission load profiles.
- The emission load profile consist of data collection phase, followed by the data pre-processing phase, the modelling phase and the data clustering phase.

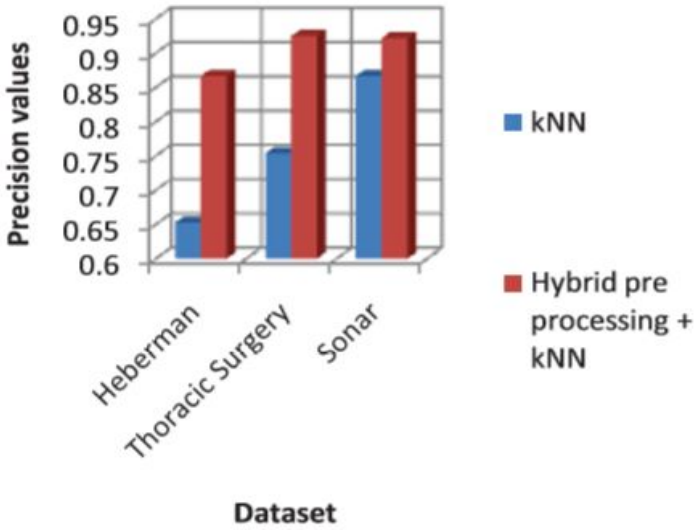


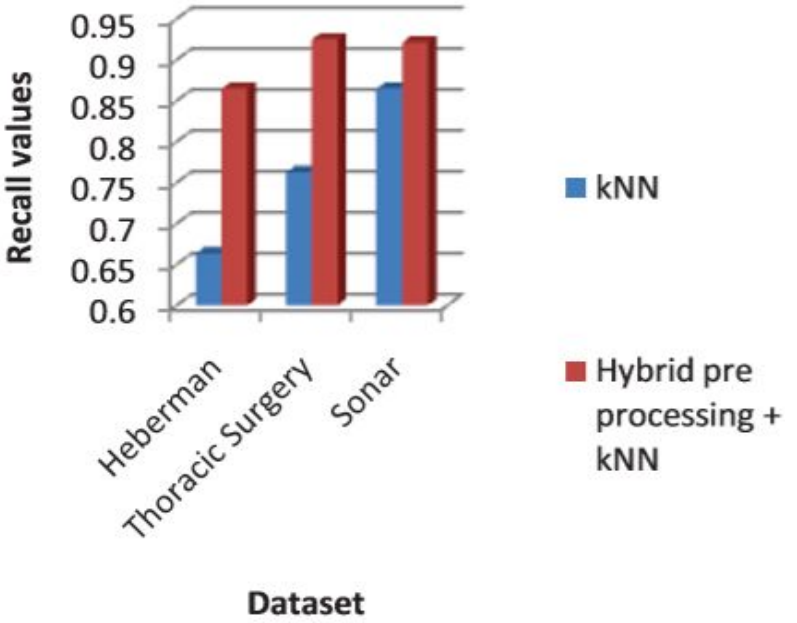
System Architecture

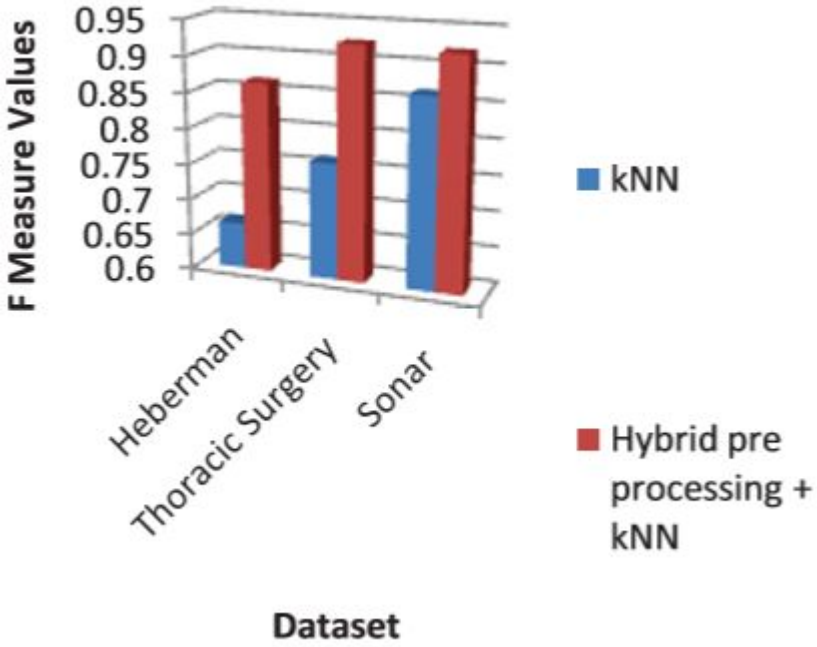


RESULT ANALYSIS

- First the dataset where executed using the existing kNN without pre-processing, Secondly the dataset was executed on the proposed hybrid technique. The results were compared in terms of performance evaluation criteria such as precision, Recall and FMeasure.




Recall values
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Heberman
Thoracic Surgery
Sonar
Dataset
kNN
Hybrid pre processing + kNN


F Measure Values
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
Heberman
Thoracic Surgery
Sonar
Dataset
kNN
Hybrid pre processing + kNN

CONCLUSION

- Data pre-processing helps in removing reductant data and hence it improves quality of the output.
- Hybrid technique handles both imbalanced data and outliers which is composed of sampling technique and a statistical technique- Interquartile range (IQR) to detect the outliers.
- Regional Emission Load Profile helps to know the amount of energy that can be conversed .

References

[1].Ashish Juneja, Nripendra Narayan Das,Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing.

[2].Preeti Nair,Indu Kashyap,Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for kNN Classifier,2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing.

[3].Angreine Kewo,Pinrolinvic Manembu,Per Sieverts Nielsen,Data Pre-processing Techniques in the Regional Emission's Load Profiles Case,2019 6th International Conference on Control, Decision and Information Technologies.

THANK YOU