Classification Of Outliers For Predicting The Heart Disease Using Distributed Data Mining With Ai

Dr. P. Ajitha

Abstract: Artificial intelligence is used for training the data to automatically predict the heart occurrence using heuristics approach. Outliers reduce the accuracy, classifying and removing it improves in predicting the heart disease. Distributed data mining helps to collect the data from various different sources to predict the heart occurrence based on the incoming data. Proposed methodology in this paper provides the heuristics approach for the faster classification and accuracy in the prediction. Existing methodologies does not use heuristics approach. Ensemble of AI and heuristics provides better approach for identifying the heart disease occurrences.

Index Terms: Artificial Intelligence, Classification, distributed data mining, heart , heuristics, Outliers, Prediction, Support Vector Machine.

1. INTRODUCTION

Distributed Data Mining is to analyze, classify and predict the data in various different sources. Distributed Data Mining handles of large volumes of data from various sources. When the nature of data is big, there is need to preprocess the data is an important one. For handling the big data preprocessing is a necessitated one. Dimensionality Reduction is another way to reduce the size of data without having major mishap on the true or essential data. Principal Component Analysis is one of the techniques to handle big data in an environment for the distributed data. When the data size is substantial, there will be group of instances which may deviate unusually from the normal or existing data and identifying that small group of instances is the goal of outlier detection. Regardless of the paucity of the deviated data, its presence may make difference to the solution model such as the distribution or principal directions of the data. Rapidly growing gap between the amount of collected data and data processing capabilities of conventional computers are very high. According to the Moore's Law, the processing power of an "average computer" doubles every 18 months, while, according to Lyman and Varian from Berkeley, the amount of stored data doubles every 12 months. In addition to this growing gap, there is an increasing need to analyze the data more quickly, more precisely, and more "intelligently". In addition to the traditional data mining tasks: classification, regression and clustering, some new challenges emerged, which require completely new algorithms for latest analysis and growing power of data. To cope with this overwhelming data flow, several frameworks for distributed data mining, together with specialized data mining algorithms, have been invented, e.g., Hadoop and MapReduce, Spark, DASK.

- *Dr. Ajitha, is currently working as Associate Professor, KG College of Arts and Science, Coimbatore, Tamilnadu, India. Email-ajitha.p@kgcas.com*

2. LITERATURE SURVEY

Distributed Communication Decision Tree Algorithm for Disseminated and Heterogeneous Environment [1] discusses the collection of data from the distributed and heterogeneous environment. The type of data it utilized was of any type or in any nature to smoothen the processing easier. Multiple sources of data collection is the base essence of distributed data mining. But all the type of data it considered was only homogeneous. This paper is one of type which dealt with heterogeneous nature. All the existing methodologies including this deals with data without outliers consideration. Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data considers and discusses the importance of data mining in healthcare for improving the medical research. Privacy issues during the collaborative data mining for medical research have been discussed.[7],[2].To solve this, an efficient approach for privacy preserving association rule mining on vertically partition healthcare data. The theoretical and practical analysis of proposed algorithm are also discussed here. Further, proposed approach can also be applied for other applications (e.g. Correlation between heart disease and food habit of patients). Big data in healthcare: management, analysis and future prospects[3] discuss the Big data analytics leverage the gap within structured and unstructured data sources. The shift to an integrated data environment is a well-known hurdle to overcome. Interesting enough, the principle of big data heavily relies on the idea of the more the information, the more insights one can gain from this information and can make predictions for future events. The exponential growth of medical data from various domains has forced computational experts to design innovative strategies to analyze and interpret such enormous amount of data within a given timeframe. The integration of computational systems for signal processing from both research and practicing medical professionals has witnessed growth. The continuous rise in available genomic data including inherent hidden errors from experiment and analytical practices need further attention. However, there are opportunities in each step of this extensive process to introduce systemic improvements within the healthcare research. High volume of medical data collected across heterogeneous platforms has put a challenge to data scientists for careful integration and implementation. The birth and integration of big data within the past few years has brought substantial advancements in the health care sector ranging

from medical data management to drug discovery programs for complex human diseases including cancer and neurodegenerative disorders. Big data analytics will march towards a predictive system. This would mean prediction of futuristic outcomes in an individual's health state based on current or existing data (such as EHR-based and Omics-based). Similarly, it can also be presumed that structured information obtained from a certain geography might lead to generation of population health information. Taken together, big data will facilitate healthcare by introducing prediction of epidemics (in relation to population health), providing early warnings of disease conditions, and helping in the discovery of novel biomarkers and intelligent therapeutic intervention strategies for an improved quality of life [5]. An efficient algorithm for detecting outliers in a distributed environment using minimal in-frequent item set pattern mining discussed the prediction of outliers in the healthcare data. The proposed methodology in this paper satisfies the accuracy in prediction. Ensemble of Artificial intelligence and heuristics proposed here provides a better classification accuracy.

3. PROPOSED METHODOLOGY:

Patient's data of their health is very much vital for predicting the diseases. As the data are vast and sometimes not preprocessed properly involves prediction of it may not be accurate. An Artificial based heurai outlier detection algorithm is proposed.

Algorithm of outlier detection: aiheuriout
Input : an dataset
Output : predicting heart disease occurrence
Input the data d[i]
a[i]=attributes list of the data
Scan the dataset for missing values
For every row or column with data apply outlier
outlier = $\approx$a[i] $\sum_n^i \aleph$
train the data with each and every data
find outliers for the data
remove the outliers for better classification
classify it
for each search of data apply heuristics
based on the heuristics results amplify the data
apply ai base heuri for incoming data
train and retrain data based on the outcome
classify based on it
if new data occurs reclassify and apply heuri
search the heuristics of best match
new results retrain
predict based on outcome of classify
accuracy calculation based on classification
ensemble the methods and predict

The algorithm is used to process the dataset for health care. Here, heart data was considered for this purpose. When a data is input to the process, basically all the type of data need to analyzed. Data obtained may be of any type of nature and which may or may not follow a pattern. So the analysis or processing of the take for classification and prediction involves more time. To reduce the time taken for processing as well as prediction of the heart rate change and impact, outliers are used. Outliers are the method of preprocessing to reduce the anomalies in data. Proposed algorithm detects outliers as a first step. After processing the data for removing outliers then heuristics is applied. Heuristics applied are based on the data it is considered. In the heart data, the heuristics . Artificial Intelligence is used to train the data automatically so that whenever the incoming data does not falls into the already existing heuristics, it will be automatically apply new heuristics and will reapply the new set. Data are removed based on the outliers. Predicting and classification involves various parameters for training the data. Artificial intelligence along with the heuristics will pave way for the closely related predicting accurately.

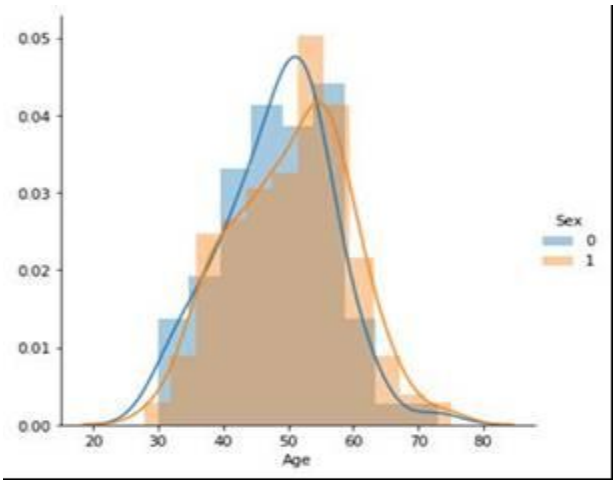

Fig 1 : *age and gender wise heart data*

Figure 1, describes the age distribution of heart data. Depicting the heart data with the chest pain type . Reason for using this data is to correlate the age to the chest pain type. Gender is mentioned as 0 and 1 as it distributes data from o to 1. Based on the age and gender the heart disease prediction are shown in the figure 1. Above figure depicts the data over the age and sex as criteria for the occurrence of the heart diseases.

Fig 2 : *Age wise heart analysis with cp .*

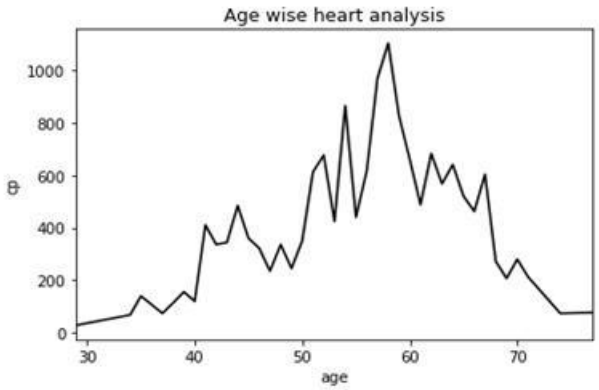

Figure 2 depicts the age wise heart analysis with chest pain type. Chest pain type differs based on the gender and age. Depicting the pain type with cholesterol level will also indicate and predict the presence of the heart diseases. The higher spikes displays the level of cholesterol and the pain type based on the age. There are various parameters for heart diseases classification and prediction.

Based on the various parameters and data, classifying

involves more time and effort. To overcome this artificial intelligence is used so that, untrained and new data can be adopted automatically.

Heuristics may vary for the type of diseases , but data considered is for heart data. Multi heuristics are applied for instant decision by the doctors by applying aiheuriout.

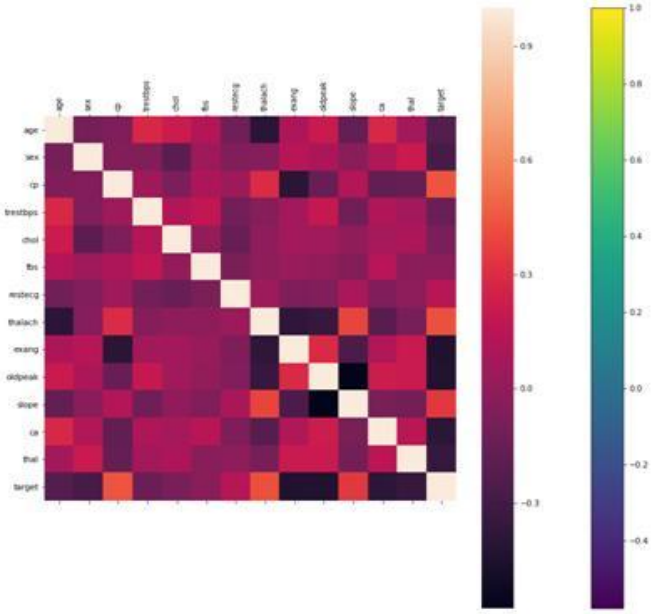

Fig 3 : *correlation matrix for heart data with all parameters.*

Figure 3, depicts the correlation matrix for data with all the parameters. This figure represents how the data are closely related and categorized based on the age for heart parameters. After, applying the algorithm the table 1, shows the distribution of the data. Parameters of the data are shown in the correlation matrix to display the correlated values for heart data. All parameters are distributed and by applying the algorithm of finding outliers. Distribution of data with heart diseases parameters are shown in the below table.

Table 1:

Distribution of Data
[[0.9521966 0.68100522 1.97312292 ... 0.71442887 2.14887271 0.91452919]
[1.91531289 0.68100522 1.00257707 ... 0.71442887 0.51292188 0.91452919]
[1.47415758 1.46841752 0.03203122 ... 0.71442887 0.51292188 0.91452919]
...
[1.50364073 0.68100522 0.93851463 ... 1.24459328 1.12302895 1.09345881]
[0.29046364 0.68100522 0.93851463 ... 0.26508221 1.12302895 1.09345881]
[0.29046364 1.46841752 0.03203122 ... 0.26508221 0.51292188 1.09345881]]

After implementation of the aiheuriout, the classification of data are done. So outliers are find to improve the prediction rate accurately.

Table 2

Outliers Criteria
(array([0, 0, 0, ..., 302, 302, 302]), array([2, 5, 6, ..., 6, 7, 13]))

Table 2, provides outliers criteria application with the algorithmic implementation. This outliers are the application of removing outliers based on heuristics applications. The proposed methodology are of novel in a way as the heuristics that applied can pave way for the decision making accurately. Existing methodologies [3]does not discuss applying for the particular diseases. It only concentrates on the health care analytics and processing of the medical information in the purely analytical way.

4. RESULTS AND DISCUSSIONS

After applying the aiheuriout, the outliers are removed properly. Based on the outlier criteria removal was done. Existing papers concentrates on outlier removal without criteria [6], but only based on Principal Component Analysis based on the dimension reduction. This may also results to the inefficient prediction.

Table 3:

Distribution of Data
[[0.9521966 0.68100522 1.97312292 ... 0.71442887 2.14887271 0.91452919]
[1.91531289 0.68100522 1.00257707 ... 0.71442887 0.51292188 0.91452919]
[1.47415758 1.46841752 0.03203122 ... 0.71442887 0.51292188 0.91452919]
...
[1.50364073 0.68100522 0.93851463 ... 1.24459328 1.12302895 1.09345881]
[0.29046364 0.68100522 0.93851463 ... 0.26508221 1.12302895 1.09345881]
[0.29046364 1.46841752 0.03203122 ... 0.26508221 0.51292188 1.09345881]]

Above table shows the distribution of the data for entire dataset of healthcare in heart diseases but with criteria application. Table 3, represents the entire parameters of heart diseases in the distribution of data with Z threshold criteria.

Table 4

Outliers Criteria
(array([28, 48, 85, 92, 158, 163, 164, 204, 220, 221, 223, 246, 248,251, 272, 281]),
array([4, 12, 4, 11, 11, 11, 11, 9, 4, 9, 3, 4, 3, 11, 7, 12]))

table 4, depicts the outliers criteria with Z threshold level an normal level of outliers removal. On comparison of table 2 and 4, both are outliers criteria but table 4 is general level criteria as it shows more outliers .table 2, shows exact removal of outliers. After, outliers removal and criteria application the algorithm implementation and efficient classification of the outliers in various data that are mined from the various and different sources of data. Distributed data are collected for parallelized and different locations which may be of homogeneous or heterogeneous nature. Removing null values is another way for efficient classification which further can be applied for training of data to use in AI trained model.

Table 6

```
fbs        403
oldpeak    421
trestbps   424
thalach    424
exang      424
chol       431
restecg    455
ID         457
Age        457
Sex        457
cp         457
num        457
Place      457
dtype: int64
(457, 11)
(342, 11)
```

```
[[2.45021366 0.53452248 1.2332679  ... 0.76856859
0.70952448 0.8786471 ]
 [2.33864755 0.53452248 1.2332679  ... 0.76856859
0.70952448 0.8786471 ]
 [2.22708144 1.87082869 2.29836291 ... 0.76856859
0.70952448 0.8786471 ]
 ...
 [1.34303417 0.53452248 0.89692211 ... 1.30112005
0.70952448 1.13811336]
 [0.45050527 1.87082869 0.89692211 ... 0.76856859
0.70952448 1.13811336]
 [1.34303417 0.53452248 1.2332679  ... 1.30112005
0.70952448 1.13811336]]
(335, 11)
```

Table 6, depicts the outliers classification and aiheuri out represents the implementation of classifying data and predicting the heart diseases with heuristics applying.

Table 7

```
Accuracy using aiheuriout: 91.0
Time taken using Support Vector Machine:
0.017000913619995117
```

After implementation of the aiheuriout the ensemble artificial intelligence and heuristics accuracy is of 91 % with outliers criteria removal and prediction.

Based on age and num

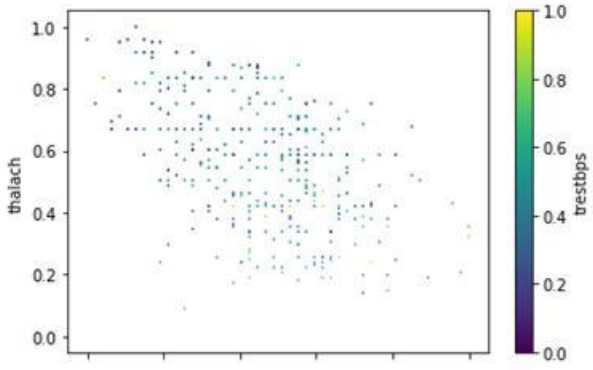

Fig: *4 Data after outliers' removal and AI application*

Figure 4, shows the data after outliers' removal and AI algorithm to predict the accuracy of implementing the algorithm

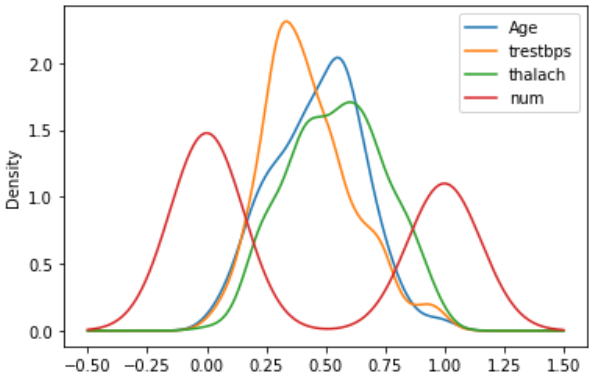

Fig 5 *prediction of heart disease*

Figure 5, mentions the prediction of heart disease with the parameters and removing outliers. Density mentions the application of outliers in parameters based on the age with rest in the heart of abs.

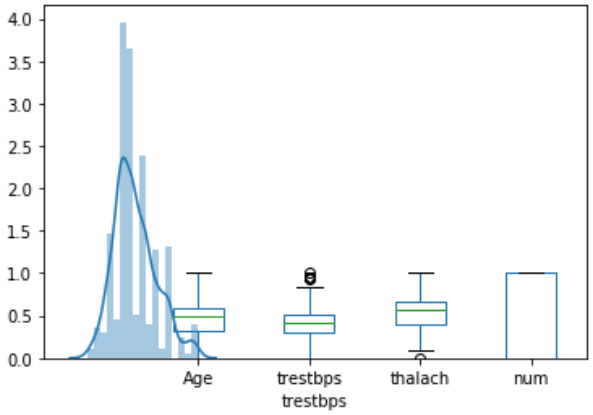

Fig 6: *AI trained to predict heart disease*

Figure 6, depicts the training of the data with using aiheuri out in prediction and classification of the data. Finally, accuracy classification of the data is much better compared to the existing methodologies. AI is automatically trained for the new data category . Unsupervised learning can be applied for the data with distributed data.

5. CONCLUSIONS AND FUTURE WORKS

The proposed methodology, accurately classifies the heart diseases. Finally, ensemble and heuristics are applied for the unsupervised data to provide good classification of the outliers and prediction. In this paper, only accuracy is considered but the level of performance are not measured in terms of higher level of prediction and classification. Future works of it may extend to the performance metrics of memory and time complexity. Heart diseases and other parameters are varying. So, a generalized level of parameterized and outliers may be devised so that robust outlier detection can be done as future work.

REFERENCES:

[1] Chandra & Ajitha, P.(2011). Distributed Communication Decision Tree Algorithm for Disseminated and Heterogeneous Environment.

Advanced Materials Research. 403-408. 1002-1007. 10.4028/www.scientific.net/AMR.403-408.1002.

[2] Nikunj Domadiya, Udai Pratap Rao, “Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data”, Sardar Vallabhbhai National Institute of Technology, Surat.

[3] Dash, S., Shakyawar, S.K., Sharma, M. et al. “Big data in healthcare: management, analysis and future prospects”. Journal of Big Data ,6, 54 (2019) doi:10.1186/s40537-019-0217-0.

[4] Chandra, P, Ajitha. (2011). “Map Reduce for DC4.5 and Ensemble Learning In Distributed Data Mining”. International Journal of Computer Science and Information Security.

[5] Abedjan Z. et al. (2019) Data Science in Healthcare: Benefits, Challenges and Opportunities. In: Consoli S., Reforgiato Recupero D., Petković M. (eds) Data Science for Healthcare. Springer, Cham

[6] Dr.E.Chandra, Ajitha.P, “A Dimensionreduct-Random Sampling Algorithm For Outlier’s Detection In A Distributed Environment”, pp. 63-73, Prapti, e-journal. ISSN: 2456-8708, 2017 .

[7] Nikunj H. Domadiya, Arpesh Kumar, Udai Pratap Rao, “Improving Healthcare using Privacy Preserving Association Rule Mining in Distributed Healthcare Data”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-4, April 2019

[8] Chandran, C.R. & Padmanabhan, A. (2016). An efficient algorithm for detecting outliers in a distributed environment using minimal in-frequent item set pattern mining. 7. 22