MRCS: Map Reduce based Algorithm for Identifying Important Features from Big Data using Chi-Square Test

Chandrashekar D. K., Srikantaiah K. C., Venugopal K. R.

Abstract*: In recent trend, big data analytics is a hot research topic for analyzing data for the business purposes, in which extraction of the important features from high volume of data is a hindrance job. In the current system, there are various methods available to extract the important feature, but it is not accurate in extraction of important features. To overcome this problem, in this paper, we have proposed a model called Map- Reduce based Chi-Square (MRCS) for feature selection. Next, the data preprocessing techniques and machine learning algorithms are used to generate business intelligence rules. The experimental results show that our proposed algorithm takes less execution time.

***Keywords:* Big Data, Business Intelligence Rules, Chi-Square, Feature Selection, Map-Reduce.**

I. INTRODUCTION

Big Data name itself represents large collection of data from various sources. For example: Health Care, Revenue (Bhoomi Project), Social Media, Banking Sector, Urban local bodies, Education Sector, from all these sectors the high amount of data is generated. Big data is an described around 4 V's: Volume, Velocity, Variety and Veracity, has attracted many interests in fixing social and financial problems, with anticipation of efficient groups and decision-making, Pre-processing this huge volume of data is a tedious job and extracting the important features is a meticulous job. The data extraction and selection of features are great extending towards research activity. In existing method, selection of features are not accurate and it consumes more time to pre-process the entire data and here, there are so many challenges in selection of features like data integration, data preprocessing. Hence, it requires effective way of selecting features using Machine Learning Techniques.

To select the important features from the given large data set Chi-Square test has been used. First, the Observed value is obtained from the data set and Expected Value is calculated and Chi-Square value is computed. The Chi-Square value is compared with the level of Significance 'α '. If the Chi-Square value is less than the level of Significance 'α ', then that particular features are selected. To do this entire process the Chi-Square test takes more time. To overcome this, we have proposed Map Reduce based Chi-Square method.

Map Reduce is a programming model to deal with the big data. Map Reduce consist of two phases: Map and Reduce phase. In Map Phase the input dataset is processed and producing some intermediate results on part of data, then the reducer phase combines the mapper results and forms the final output.

In this paper, we have proposed Map Reduce based Chi-Square (MRCS) for selecting important features using machine learning techniques. To Design a model by collecting a data from various sectors and knowledge discovery of the domain from where the data is generated. Data preprocessing techniques can be used on selected features and generates business intelligence rules.

The rest of the paper is organized as follows, Related Works are explained in Section 2, and Problem is defined in Section 3, and system architecture (MRCS) is elaborated in section 4, the proposed model is explained in Section 5, Experimental Results shown in Section 6, Conclusion and future work described in Section 7.

II. RELATED WORKS

Vinod *et al.,* [1] proposed a Highly Correlated Feature Set Selection (HCFS) algorithm for classifying big data of patient's records by using hierarchical learning approach for improving the accuracy of classifications.

Kleerekoper *et al.,* [2] designed a structure to remove the noise from selected features by eliminating all irrelevant data and redundancy data based on Manchester Analytics Toolkit (MAST) and information theory. It encounters all the zero valued data in the structure of array for reducing the memory usage and execution time.

GirijaAttigeri *et al.,* [3] proposed a clustering and classification algorithm for identifying a fraud in financial system. The main goal is to dimensionality reduction for selection and extraction of features in huge volume of data. By doing this, the execution time is reduced the accuracy is improved.

Jundong Li *et al.,* [4] mentioned the challenges of features selection in big data by considering the various types of data like structured featured data, link data, multisource, Multi view data, streaming data and features, stability and scalability.

Revised Manuscript Received on December 15, 2019.

***** Correspondence Autor

Chandrashekar D K, Assistant Professor, Department of Computer Science and Engineering, SJB Institute of Technology, Bangalore, Karnataka, India. Email: dkchandrashekar28@gmail.com

Srikantaiah K C, Professor, Department of Computer Science and Engineering, SJB Institute of Technology, Bangalore,Karnataka, India. Email: srikantaiahkc@gmail.com

Venugopal K R, Vice-Chancellor of Bangalore University, Bangalore, Karnataka,India. Email: chandu_dchally@yahoo.com

Liang Zhao *et al.*, [5] addressed the challenges of efficient analyzing of the high dimensional of economic data based on distributed selection of feature. The goal is to extract the features for generating rules from economic big data. and for analyzing the accurate results for big data.

Conor Fahy *et al.*, [6] designed an independent dynamic feature mask algorithm for streaming high dimensional data, which addresses challenges of selects features dynamically and removing redundant data using density based clustering algorithm.

Wang, *et al.*, [7] proposed a framework for feature selection which globally minimizes the redundant data and maximizes the score ranking and also introduced an efficient global unsupervised and descriptive local selection of features.

Çatalkaya *et al.*, [8] developed software architecture to combine different types of feature selection by using chi-square method to produce good results for banking data.

Ogudo *et al.*, [9] designed a model called SQM to traverse the benefits of In-Memory, processing of big data and very less cost business Intelligence tools to represent a good Service Quality Management.

Ping Yu *et al.*, [10] proposed a features selection algorithm for agriculture big data using cloud services as platform. Transformation of data from data with primitive rough extraction features to data with high-level semantic features is of great significance for target task learning.

III. PROBLEM DEFINITION

Consider an integrated data set *'D'*, with *n* attributes or Features *i.e*, '*F*'= {f_1, *f2, f3*....f_n), and *m* tuples or rows, our main objective is to select important features from *F* using Map Reduce based Chi-Square Test.

IV. MAP REDUCE BASED CHI-SQUARE

A. System Architecture

The system architecture consists of the following components: (i) Data integration (ii) Feature Selection (iii) Data Cleaning (iv) Data Analysis (v) Business Intelligence Rules (vi) Cloud as appear in Figure.1

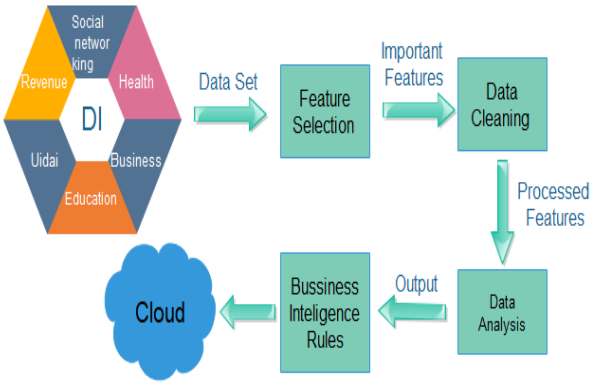


Fig. 1.System Architecture.

Data Integration: Data generation from different sectors like UIDAI, social networking, Health, Business, Education and revenue etc, as show

n in Figure1. Each data will be in different style , the data type and formatting it vary from one style to another style combine all the different form of data is called data integration.

Feature Selection: Feature selection is used to take out the unrelated and repeated data from the data set and returns important features. Feature selection improves the accuracy level and dimensionality reduction in data, the extracted features is given to the cleaning process.

Data Cleaning: Data cleaning is done after the feature selection to reduce time because time taken to clean the data set is more compared to cleaning the feature selection. The data cleaning removes the irrelevant and inaccurate data the purpose of cleaning the feature selection for accurate results. The process is critical and emphasized because wrong data can drive business to wrong direction and decision to perform in huge data.

Data analytics: Machine learning algorithms are applied on selected important features which discover the patterns, decision rules, prediction model and very useful information.

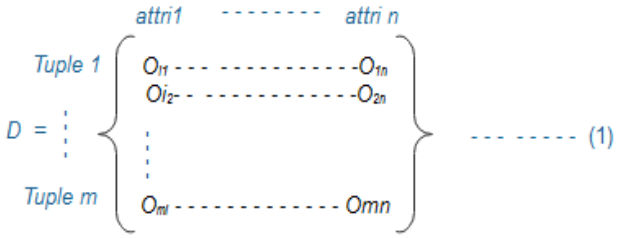
Business Intelligence Rules: Finally, the generated Business Intelligence rules are used to take decisions for business purposes.

Cloud: The generated Business intelligence rules are stored in cloud. and it is used for taking decisions in business.

V. PROPOSED METHOD

A. Chi-Square

Chi-square test is based on supervised learning, which works on test data, trained data and target value. The Chi-Square test is performed on each feature in the trained data set as shown in equation 1. Training data set is a m *X n* matrix, where m represents number of *tuples* or rows and *n represents* number of attributes, O_{ij} represents observed value for j^{th} attribute in i^{th} tuple.



By using equation 1 we need to find out the Observed value *Oi* for each feature the null hypothesis is a prime concern to know how data is distributed.

The null hypothesis for each Chi-Square test can be stated as

$H0: O_i = E$

$H1: O_i \neq E$

After obtaining the Observed Value O_i and find the Expected value *E* for each feature in the data set using the equation 2.

$$E = \frac{N}{n} \quad \text{.....} \quad (2)$$

N is Total number of Values and *n* is Number of Features in the dataset.

The degree of freedom for the chi-square distribution is calculated based on the measure of the contingency table using equation 3.

Degree_ of_ freedom = *(m - 1) * (n – 1)……. (3)*
After obtaining the observed value O_i and Expected value E for each attribute i and degree_of_freedom. The chi-square value for each attribute is calculated for Selecting an important features by using equation 4.

$$\chi_i^2 = \chi_i^2 + \frac{O_j - E}{E} \quad \text{..... (4)}$$

Chi-Square test results are interpreted in the view of chi-square distribution with the required number of degrees of freedom and Compares the obtained chi-square value with minimum hypothesis of independence critical value α=0.05

If $\chi_i^2 < \alpha$ which accepts null hypothesis (*H0*), it is completely independent features does not depend on any other features by using the following equation 6.

$$\chi_i^2 < \chi_{\alpha=0.05}^2 \ldots\ldots\ldots\ldots \quad (5)$$

If the chi –Square value of any feature is less then minimum hypothesis of independence critical value then that particular feature is selected as shown in Table 1. The detailed algorithm for selecting important features using Map Reduce based Chi-Square (MRCS) is explained in Algorithm 1.

Table-1 Identified important features using Chi-square

Method	Dataset	No of attributes	Important attributes
Chi-Square	Economic	28	6

Algprithm1: Map Reduce based Chi-Square (MRCS)

MRCS (D, α, *featurelist*)
Procedure: To generate the important features
Input: *D* - Economic data set, α- Level_ of _Significance
featurelist- attributes of D, N- total number of Values in the data set, n- number of features.
Output: *SelectedFeatures: Set of Important Attributes*

Begin
SelectedFeatures = ϕ, α=0.05, $E_i = \frac{N}{n}$
For each feature *i* ∈ *Featurelist* do
Chi_Mapper (*i*)
End For
Chi_Mapper (*i*)
{
For each tuple *j* in *D* do

$$\chi_i^2 = \chi_i^2 + \frac{O_j - E}{E}$$

End For
}

Chi_ Reducer (i)
{
For each feature *i* ∈ *Featurelist* do
If (Chi-square[i] < α)
SelectedFeatures = U featurlist [i]
}
End

VI. EXPERIMENTAL RESULTS AND DETAILS

A. Experimental Setup

The MRCS model is implemented using Apache Spark 2.4.3 version on (5.5 LTS) Data bricks. Spark is designed based on the language python version-3. The system is configured as 8GB Ram, i5 processor, 1TB Hard disk, 10 CPU Cycles. The world economic data set of the year 2018 and 28 attributes with .csv format is considered.

B. Feature Selection

The Figure 2 shows that X axis shows the features and Y-axis shows the importance of features. There are 28 attributes are given as an input When comparing the Chi-Square value χ_i^2 with Level of Significance **α** only 6 features are identified as an important. Such as cost of business start-up procedures, other taxes payable by business, time required to get electricity, time required to register property, time required to start a business and time to resolve insolvency as shown in Figure 3.

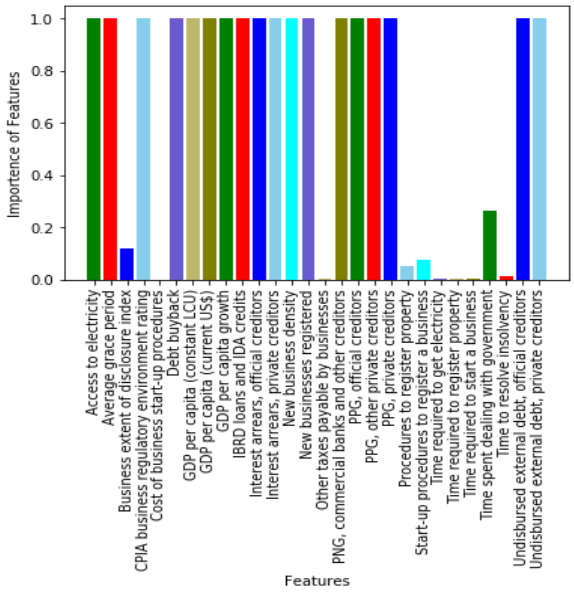


Fig. 2.Identifying Features using Map Reduced Chi-Square.

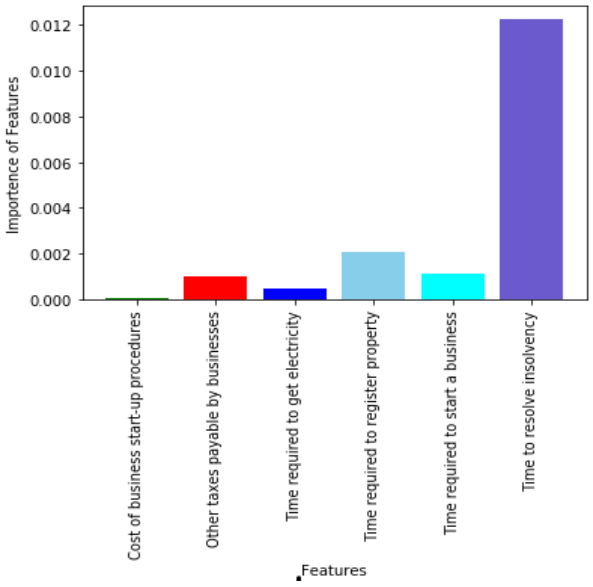


Fig. 3.Identifying important Features using Map Reduced Chi-Square..

C. Performance Evaluation

Table 2. Shows the Results of the proposed system Map Reduce Chi-Square with the Chi-Square. Map Reduce chi-square takes less execution time for identifying important features from the given data set. Because Chi-Square value for each attribute is calculated using mapper function in parallel.

Table 2: Execution time comparison between chi-square and Map Reduce Chi-Square for identifying important features

Data set			Time taken (Sec)	
No of records	No of features	Important features	Chi-Square	Map Reduce Chi-Square
100	28	6	100	10
1000	300	120	200	25
10000	5800	2500	500	55
100000	35000	15000	1000	100

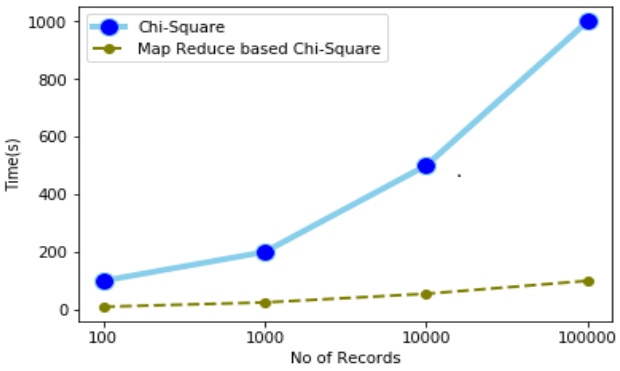


Fig. 4.Execution time comparison between chi-square and MR chi-square.

X-axis shows number of records and y-axis shows time taken for execution in seconds. The Figure 4 shows that time taken to extract the important features with or without Map Reduce. The Map Reduce based chi-square performs the good results when compare with chi-square in seconds.

VII. CONCLUSION

In this paper, The Map Reduce based Chi-Square (MRCS) model is designed to Select Important features for business. MRCS model performs good results in execution time compared to the existing method, the paper concludes that the execution time for extraction of feature selections is reduced. Further, the data pre-processing techniques can be applied to remove all irrelevant and redundancy on only selected important features and also business intelligence rules are generated using machine learning techniques on preprocessed important features for business purpose.

REFERENCES

1. D. F. Vinod and V. Vasudevan, "A filter based feature set selection approach for big data classification of patient records," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 3684-3687.
2. A. Kleerekoper, M. Pappas, A. Pocock, G. Brown and M. Lujan, "A calable implementation of information theoretic feature selection for high dimensional data," *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, 2015, pp. 339-346.
3. Attigeri, Girija, MM Manohara Pai, and Radhika M. Pai. "Analysis of feature selection and extraction algorithm for loan data: A big data approach." *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017.
4. Li, Jundong, and Huan Liu. "Challenges of feature selection for big data analytics." *IEEE Intelligent Systems* 32.2 (2017): 9-15.
5. Zhao, L., Chen, Z., Hu, Y., Min, G., & Jiang, Z. (2016). Distributed feature selection for efficient economic big data analysis. *IEEE Transactions on Big Data*, *4*(2), 164-176.
6. Fahy, Conor, and Shengxiang Yang. "Dynamic Feature Selection for Clustering High Dimensional Data Streams." *IEEE Access* 7 (2019): 127128-127140.
7. Wang, De, Feiping Nie, and Heng Huang. "Feature selection via global redundancy minimization." *IEEE transactions on Knowledge and data engineering* 27.10 (2015): 2743-2755.
8. Mehmet *Burak*, *Oya Kal*. "Data Feature selection methods on Distributed data processing platforms." *3rd international conference on computer science and engineering.* (IEEE 2018): 133-138.
9. Ogudo, Kingsley A., and Dahj Muwawa Jean Nestor. "Modeling of an efficient low cost, tree based data service quality management for mobile operators using in-memory big data processing and business intelligence use cases." *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. IEEE, 2018.
10. Yu, Ping, and Hui Yan. "Study on Feature Selection and Feature Deep Learning Model For Big Data." *2018 3rd International Conference on Smart City and Systems Engineering (ICSCSE)*. IEEE, 2018

AUTHORS PROFILE

Chandrashekar D K is currently working as Assistant Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India. And pursuing the Ph.D degree in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, under Visvesvaraya Technological University Belgavi, India. He obtained his B.E degree in 2009 and M.Tech degree in 2014 from Visvesvaraya Technological University Belgavi, India. His research interest is in Data Mining, Big Data Analytics and Cloud Computing.

Srikantaiah K C is currently working as Professor in the Department of Computer Science and Engineering at S J B Institute of Technology, Bangalore, India. He obtained his B.E from Bangalore Institute of Technology, M.E from University Visvesvaraya College of Engineering, Bangalore in 2002 and Ph.D degree in Computer Science and Engineering from Bangalore University, Bangalore, in the year 2014. He is guiding five Ph.D students in VTU. During his 15 years of service, he has 20 research papers to his credit. He has authored a book on Web Mining Algorithms. He has awarded best paper presentation award in the conference ICIP 2011 and his name is listed in Marquis who is who in the World 2014, 2015, 2016. His research interest is in Data Mining, Web Mining, Big Data Analytics, Cloud Analytics and Semantic Web.

Venugopal K R is currently working as a Vice-Chancellor of Bangalore University. He obtained his Bachelor of Engineering from University Visvesvaraya college of Engineering. He received his master's degree in computer science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras.

He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 57 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc., He has filed 101 patents. During his three decades of service at UVCE he has over 550 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed.