Feature Selection Method based on Fisher's Exact Test for Agricultural Data

S. Rajeswari, K. Suthendran

Abstract*: This paper is aimed to analyze the feature selection process based on different statistical methods viz., Correlation, Gain Ratio, Information gain, OneR, Chi-square MapReduce model, Fisher's exact test for agricultural data. During the recent past, Fishers exact test was commonly used for feature selection process. However, it supports only for small data set. To handle large data set, the Chi square, one of the most popular statistical methods is used. But, it also finds irrelevant data and thus resultant accuracy is not as expected. As a novelty, Fisher's exact test is combined with Map Reduce model to handle large data set. In addition, the simulation outcome proves that proposed fisher's exact test finds the significant attributes with more accurate and reduced time complexity when compared to other existing methods.

Keywords* : *Fisher's exact test, feature selection, MapReduce, chi-square test, statistical test.

I. INTRODUCTION

Data from real-world applications can be of high dimensional. Classification techniques are used to assign subjects to a specific class of a target variable. In classification problems, predictive models are developed to predict the target variable based on several input variables (features). Features, which are also referred to as attributes, are independent variables. This is particularly true for the applications in all fields of text categorization, bio-molecular analysis, scene classification, remote sensing, and medical diagnosis. With the rapid development of modern technology, tremendous new computer and internet applications have generated large amounts of data at an unprecedented speed, such as video, photo, text, voice, and data obtained from social relations and the rise of the Internet of things and cloud computing.

These data often have the characteristics of high dimensions, which poses a high challenge for data analysis and decision-making. Feature selection has been proven in both theory and practice effective in processing high-dimensional data and in enhancing learning efficiency. Feature selection is referred to the process of obtaining a subset from an original feature set according to certain feature selection criterion, which selects the relevant features of the dataset.

Methods for selecting features or attributes are useful for reducing the dimensionality of the data (called subset) by ignoring the unnecessary and not related attributes in a data set. It is categorized as filter and wrapper technique. Purely the first technique relay on the general details of the data while the wrapper technique works on the basis of the rank or weight of the features. After subset generation, for each feature, the weight or rank is calculated based on wrapper methods viz., Correlation, Gain Ratio, Information gain, OneR, Chi-square MapReduce model, Fisher's exact test [1].

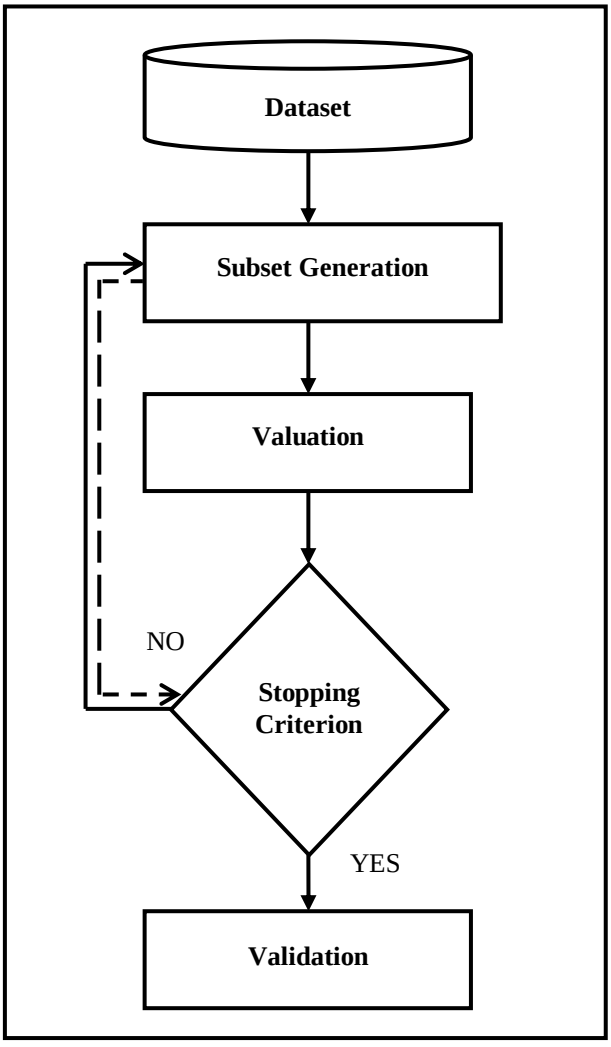

Fig. 1. Steps of Feature Selection Process.

Revised Manuscript Received on December 05, 2019.
***** Correspondence Author
S. Rajeswari*, Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnan koil - 626126, Tamilnadu, India. Email: rajeswari@klu.ac.in
K. Suthendran, Department of Information Technology, Kalasalingam Academy of Research and Education, Krishnan koil - 626126, Tamilnadu, India. Email: k.suthendran@klu.ac.in

Wrapper technique utilizes the learning algorithm to determine all subsets of identified attributes, because it always simpler than the first method, although it is expensive. Implementation of the filter method is incredibly easier, Hence, this is an excellent fit for a Big database; however, these techniques ignore the correlation between performance-affecting attributes. The attribute selection process has several advantages, such as data characteristics being understandable, reduced processing time and minimal space for the mining process, thus improving the efficiency of the algorithms [2].

Figure 1 shows the selection process of the attributes is:

- Subset Generation: Subsets of attributes are generated for evaluation on the basis of a specific search policy (It can be a null subset or Full subset or with a arbitrary subset of features)
- Valuation: Some generation criterion measures the quality of the generated subset using valuation function. The newly produced subset is evaluated and replaced with the previous best, if it is found to be better
- Stopping Criterion: The process of feature selection may continue to generate feature subsets therefore, an exit measures is essential for this selection process. For illustration, stopping the process at pre-determined number of attributes or else iterations or while adding or else deleting a feature reduces performance otherwise when obtaining an optimal subset of features
- Result Validation: Finally, the subsets are verified and validated based on datasets. In addition, the resulting subset is compared with previously identified artificial datasets, real-world datasets, or combination of both [3]

A complex classification problem involves a large number of features. The classifier will take a longer time to classify the observations when the number of features is very large. Several feature selection methods have been developed to solve classification problems. It plays a role in compressing the data processing scale, where the redundant and irrelevant features are removed.

Feature selection technique can pre-process learning algorithms, and good feature selection results can improve learning accuracy, reduce learning time, and simplify learning results. Main aim is used to propose a suitable feature selection methodology for handling large datasets with high accuracy and experimentally verify the impact of different entropy-based and statistical classifiers [4].

Compare the many statistical feature selection methods to discover the finest one for selecting relevant attributes in the agricultural data. In that, the proposed method fisher's exact test with MapReduce method gives the significant attributes also compute with less time compare than the other existing attribute selection methods.

The manuscript is prepared as follows. The section II deals with reviews of previous work about feature selection methods. Section III contains specific problems related to various ranking features and techniques for selecting features. Section IV shows the outcome of the proposed method. Final section V contains result analysis of this study, some closing statements, and issues that remain to be tackled and will be investigated in upcoming studies.

II. RELATED WORK

Variable selection (or) process of feature selection for linear relationships is more or less resolved. Unfortunately, there is often no distinction between the design for variable selection and the other selection norms. The same algorithm may results in different variable selection using dissimilar criteria and vice versa [5]. Many authors have recommended techniques to pick the required features from dataset and discussed about the algorithms for the attributes selection. Some of the research works are as follows:

El-hasnony et al. [6] analyzed five data downsizing techniques such as gain ratio, analysis of main workings, selection of relationship between the features, selection of draft set attributes, and selection of fuzzy draft features. Based on the classification accuracy, the analysis was performed and the outcome demonstrated that UCI machine learning repository's fuzzy rough feature selection exceeded other techniques used for two medical data sets, i.e. breast cancer and thoracic surgery.

Villacampa [7] considered popular feature selection methods and evaluated the performance. The classifiers based on Decision Tree, k-Nearest Neighbor and Support Vector Machines are used for classification. The results showed that all other methods for selecting features were outperformed by the Relief-F method.

Porkodi [8] analyzed five techniques of selection of features to reduce the subset's dimensionality. The outcome witnesses that the Random Forest algorithm performs better than other presented algorithms. Georg Rub and Rudolf Kruse proposed a novel application for the same process using forward feature selection methods for agriculture dataset. Further, this finds real time application using AI techniques [9].

Surabhi Chouhan et al. proposed a new methodology for feature selection process based on PSO-SVM .The Decision tree algorithm is deployed for classification. Mushroom and Soyabean benchmark datasets are considered. The outcome reveals that the proposed hybrid methodology performs better than other presented methodologies [10].

Parimala and Nallaswamy have investigated the Feature Selection methods using spam E-mail data set. Based on the investigation, they projected "Fselector" an innovative Feature Selection technique which supports high dimensional data. Using five Feature Selection methods viz., CFS, LDA, RF, Rpart and NB the performance was evaluated and presented. [11].

Pinar Yildirim [12] study found that feature selection methods can improve learning algorithms performance. The experiments are performed on four classification algorithms using renowned hepatitis data set and the presented algorithm is evaluated.

In that, the results classifiers of Naïve Bayes and Decision tree have produced better accuracy than the other existing methods. Rajeswari et al. proposed a chi-square MapReduce model for selecting the relevant features from agricultural dataset. This model handles large amount of agricultural data. The authors considered four feature picking methods; and the investigation was performed to identify the best method for agricultural dataset. The proposed model resulted in high accuracy and less processing time than the existing feature selection methods [13].

The above literature confirms that the majority of researchers have proposed feature selection algorithms using small dataset. They didn't handle the large amount of dataset or real time dataset. Data mining methods are not enough to examine these high size dataset. Therefore, there is a demand for big data analytics concept to handle the same. However, still there is a scope for improving the accuracy with reduced processing time. A novel feature ranking model was proposed in this work to achieve the same.

III. FEATURE SELECTION AND RANKING ALGORITHM

The machine learning literature has proposed numerous feature rankings and feature selection techniques. These feature techniques are intended to remove inappropriate or unnecessary characteristics from a vector of a given feature. To downsize the dataset, sub-set generation and sub-set evaluation is used. Using the filter method, each subset is evaluated [14].

In this paper, the various feature selection algorithms as mentioned in the above were considered to evaluate the practical effectiveness of the following ranking, commonly used methods, statistical and entropy-based, with good performance in different fields:

A. Correlation based Attribute Selection

This algorithm calculates the relationship between all characteristics and the output class and uses correlation-based heuristic evaluation function to select the suitable feature subset [15]. It measures the relationship between nominal or categorical characteristics, so the discrete values are used in numerical characteristics. Selection of the correlation feature is provided by equation 1.

$$r_{zc} = \frac{K\overline{r_{zi}}}{\sqrt{K + K(K-1)\overline{r_{ii}}}} \quad (1)$$

Where r_{zc} denotes the relationship between features and class variable, K represents the number of features, $\overline{r_{zi}}$ indicates the mean value of correlated feature-classes and $\overline{r_{ii}}$ represents the mean value of inter-correlated features [16].

B. Information gain Ratio

Information gain measures the amount of data obtained by the availability of the term in a text for class prediction. It picks the required attributes with respect to the information value obtained by generating a subgroup on the class attribute. Indexes of information conjecture are often required for the evaluation of features. The ultimate reason of this work is to dispose of unwanted or Information gain feature selection, the entropy value has also been calculated for whole data [17]. It is an algorithm for selection of features supervised, univariate, simple, powerful, symmetrical and entropy-based. The information gain for a feature X and class label Y as follows.

$$Informationgain(X,Y) = H(X) - H(X \mid Y) \quad (2)$$

Where H(X), H (X|Y) is calculated on X and Y for entropy values. X entropy can be computed as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (3)$$

X|Y entropy calculation is shown below

$$H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j)) \quad (4)$$

This method calculates the ratio separately on behalf of each and every attribute and selects 'm' as the most relevant features among 'n' features with high information gain, i.e. it considers a feature F with high information gain as the most relevant. This algorithm's main drawback is that it selects the feature that can or may not be more informative with high information gain. The Information Gain cannot handle redundant features as the features are selected in a univariate manner [18].

C. Gain Ratio based Attribute Selection

Gain Ratio is the modified version of the information gain. It considers daughter nodes wherein an attribute splits the dataset with respect to the class. This limits the fondness of having four attributes and huge possible values in the information gain method [19]. Equation 5 gives gain ratio.

$$GainRatio = \frac{InformationGain}{H(X)} \quad (5)$$

When predicting the variable Y, we normalize the gain of information by dividing it by X entropy, and vice versa. Gain ratio values are ranges between 0 and 1, due to this normalization. Gain Ratio = 1 point out that X knowledge predicts Y utterly, and there is no relationship between Y and X when it is = 0. It favors variables with least values as opposed to Information Gain.

D. OneR based Attribute Selection

OneR is a simple Holte-designed algorithm [20]. From the generated rules it selects the rule having least error for training. It divides the dissimilar values into many. Considers "missing" as a suitable value, and manages the missing values.

In addition, it creates rules for all features and selects a baseline performance as a benchmark for different training methods. This algorithm uses OneR classifier to determine the weights of the attributes. It generates a rule using the same attribute for each attribute and also measures its error rate afterwards [21].

E. Chi-square MapReduce Model

Chi-Square feature selections have two parameters viz., observed and expected frequency. It is calculated by MapReduce techniques [22]. The weights of the attributes can also be discovered. The corresponding attributes are the highest weight attributes. This method analyzes the class label. This is used for selecting the variable of the predictor. This attribute value with ' r ' and ' c ' class number is defined as

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (6)$$

Where O_{ij} is the number of 'i' value occurrences in class 'j'. E_{ij} is the number of events predictable with the value 'i' and the class 'j' [32].

From the above five feature selection methods advantages and disadvantages are as follows:

- Attributes with many distinct values, the information gain fails to accurately discriminate among the attributes
- It does not work good for attributes with large number of distinct values
- Information gain gives the result as 0 and 1 only
- Chi-square test is applied in a four fold table will not give a reliable result with one degree of freedom if the expected value in any cell is less than 5
- It does not measure the strength of association
- The statistical finding of relationship, does not indicate the cause and effect

The above disadvantages are overcomed by the fisher's exact test feature selection method.

F. Fisheer's Exact Test

Using the well-known Fisher ratio concept, the Fisher score is used to select required features and a heuristic policy is deployed to determine a score for features [23]. The advantages are

- To identify the relevant features for any specific problem
- Reduces the size of the problem and computer storage
- Reduce the computation time also to improve the quality of prediction
- To improve the classifier by removing the irrelevant features and noise

Let the average and standard deviation of the *k* -th class and *i*-th function (and μ^i, σ^i the mean and standard deviation of the entire dataset for the *i*-th function) be μ^i_k and σ^i_k. The *i*-th feature Fisher score (x^i) can be calculated as

$$F(X^i) = \frac{\sum_{k=1}^{K} N_K(\mu_k^i - \mu^i)^2}{\sum_{k=1}^{K} N_k(\sigma_k^i)^2} \qquad (7)$$

Where N_k is the number of Class C_k patterns. Since this score is independently calculated, a suboptimal set may be represented by the selected features. [24].

The above attribute selection methods are compared with higher accuracy to find the best suitable method.

IV. EXPERIMENTAL RESULTS

Selection of attributes is one among the pre-processing methods that ignores the irrelevant and redundant attributes and thus results high accuracy [25]. It does not imply the reduction of cardinality, but also the variety of attributes based on the presence or lack of interaction between the attributes and the algorithm for classification. This process is mandatory because the learning task is challenged by the high dimensionality and vast amount of data.

The irrelevant features selection processes during the learning algorithm have become computationally complex, over-fit, less understandable, and less accurate learning. Filter, wrapper, embedded and hybrid methods are categorized as attribute selection methods. This categorizes filter methods as univariate and multivariate. Univariate filter strategies ignore feature dependencies which, compared to alternative feature selection techniques, may result in the choice of unnecessary options and worst classification performance.

Figure 2 shows the proposed method workflow. To remove the unwanted information in the dataset, it has the data pre-processing steps and then selects the relevant prediction features using the attribute selection methods. Discover the finest method of selecting attributes and compare it to other four methods of selecting features.

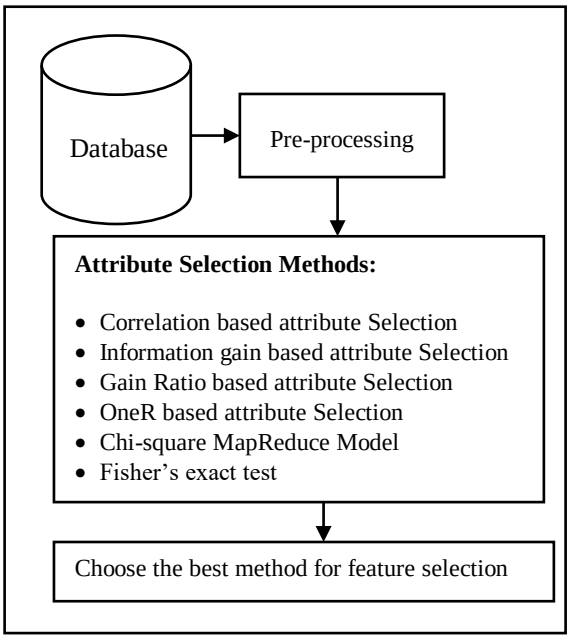

Fig. 2. Steps of Attribute Selection Process

A. Used Environment and RPackages

R is a software package and is often suitable for numerical data processing. The large amount of data is not handled by it. To integrate the R and Hadoop environment, the Rhadoop, RmR, Rhdfs packages are used. It easily processes the high volume of data. Using Rhadoop packages, the proposed approach coding was done [26].

The working environment is R in conjunction with the package kernlab [27]. It is a kernel-based learning package that offers multiple methods. The program is written in the language of R programming. Intel Core 2 Duo E7500 @ 2.93GHz with 2 GB of RAM, configuration PC was used for experiment.

B. Dataset with Data Pre-processing

The Dataset was collected from the Virudhunagar district Soil and water laboratory test center, Virudhunagar district, Tamilnadu for the 2015-2018 year data. The dataset is having soil information for 11 blocks. This dataset contains 15 attributes with 25000 total instances from the soil sample testing report. The same soil sample test results attributes are used in previous work [32].

During the pre-processing, the unwanted information in the soil data set is spotted and removed [29]. Before removing omitted values, the unprocessed dataset have the 25000 instances. In that, incompatible data, fault data, and absent values be ignored. Later than the pre-processing process the soil test report contains the 15 features with 24980 total instances for Virudhunagar district 11 Blocks. So, that the pre-processed dataset features are taken into the feature selection process.

C. Performance Measures

The accuracy calculation is a very essential component for any classification algorithm. It will testify the past active classification algorithms as “Right” or else “Wrong” [30].

The proposed method accuracy of is compared by the other existing methods. The accuracy was considered by the whole amount of all accurate predictions divided by the whole amount of the data [31]. In that whole data set, 25% of data is allotted for training phase and the remaining allotted for testing phase. The accuracy is measured by,

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN} \right) \quad (8)$$

Where, TP denotes True Positive for identified the total number of correctly positive classify occurrences, TN represents True Negative for recognized the total number of correctly negative classify occurrences, FP denotes False Positive for identified total number of incorrectly positive classify occurrences, and FN represents False Negative for identified total number of incorrectly negative classify occurrences. The performance of the algorithm is calculated using the above formula 8.

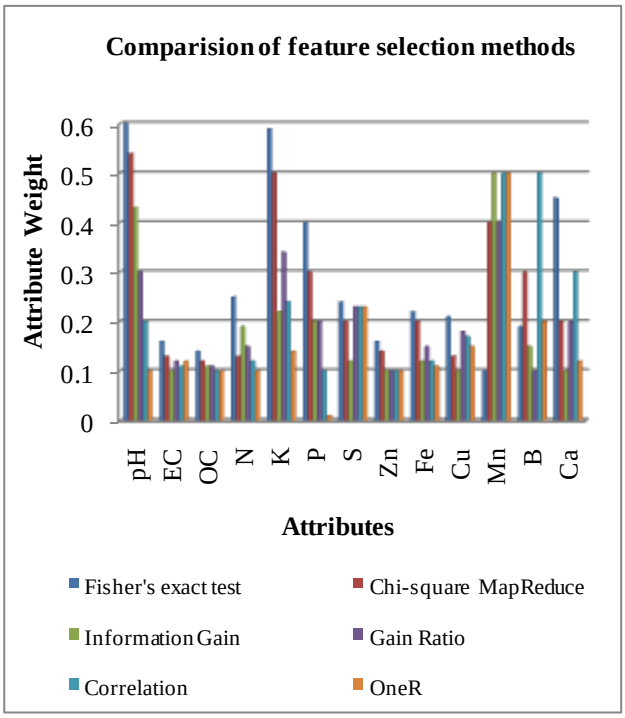

Fig. 3. Comparision of Feature selection result

Figure 3 shows the weights for the feature selection methods and is compared. In that, Fisher’s exact test gives the relevant attributes compare with the other feature selection methods are Correlation, Information gain, Gain Ratio, OneR, and Chi-square MapReduce Model. The highest weight attributes are considered as relevant attributes are pH value, Phosphorus (P), Potassium (K), Nitrogen (N), Iron (Fe), Sulphur, Boron (B), and Organic Carbon (OC) is taken into the further classification algorithms.

Table- I: Features Weight Results

Features	One R	Correlation	Gain Ratio	Information Gain	Chi-Square MapReduce	Fisher’s Exact test
pH	0.1	0.2	0.3	0	0.54	0.6
EC	0.12	0.11	0.12	1	0.13	0.16
OC	0.1	0.1	0.11	0	0.12	0.14
N	0.1	0.12	0.15	0	0.13	0.25
K	0.14	0.24	0.34	1	0.5	0.59
P	0.01	0.1	0.2	1	0.3	0.4
S	0.23	0.23	0.23	1	0.2	0.24
Zn	0.1	0.1	0.1	1	0.14	0.16
Fe	0.11	0.12	0.15	0	0.2	0.22
Cu	0.15	0.17	0.18	1	0.13	0.21
Mn	0.5	0.5	0.4	0	0.4	0.1
B	0.2	0.5	0.1	0	0.3	0.19
Ca	0.3	0.2	0.13	0	0.34	0.12

Table 1, shows the attribute weights for each feature selection methods. From the attributes weights, based on the threshold value find the highest weights features for fertility level prediction. Each and every method predicts the relevant attributes and irrelevant attributes. These pH, P, K, N, Fe, OC, B, S features are important the fertility level prediction. Soil pH is a measure of the acidity and alkalinity in soils. pH levels range from 0 to 14, with 7 being neutral, below 7 acidic and above 7 alkaline. Based on that, correctly and incorrectly features are identified.

Table –II: Comparision Results of Feature Selection Methods

Feature Selection Methods	Correctly identified Features	Incorrectly identified Features	Accuracy (%)	Time (in secs)
Correlation	5	7	80%	8.5
Information gain	6	6	85%	6.25
Gain Ratio	7	5	90 %	5.0
OneR	8	4	91 %	4.75
Chi-square MapReduce Model	9	3	97 %	3.15
Fisher's exact test	10	2	98 %	2.25

Table 2, shows the experimental outcome of six feature selection methods. In that, Fisher's exact test provides high accuracy (98%) and also less time (0.5 sec) taken to find the relevant features was compared with the other existing feature selection methods.

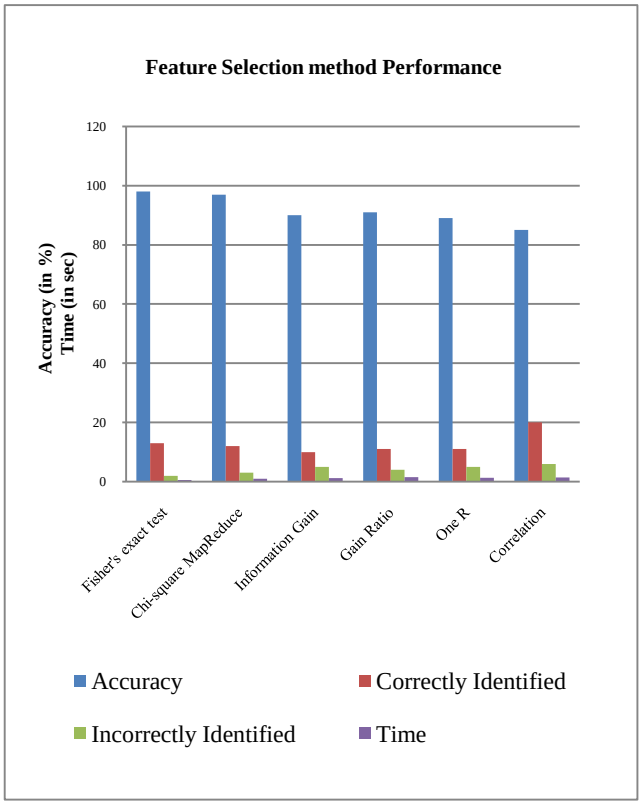

Fig 4. Feature selection performance analysis

Also Figure 4. shows the evaluation result of precision, execution time, correctly, and incorrectly identified features of proposed Fisher's exact test method with other feature selection methods. Since, the proposed method produces the relevant features from large data set for predict the farmland soil fertility level.

V. CONCLUSION AND FUTURE WORK

Feature selection plays an essential role in data mining pre-processing step and also several data mining machine learning methods are finding difficult to handle large amounts of irrelevant features. Therefore, there is a demand for suitable feature selection approach. In this paper, as a novelty, Fisher's exact test is combined with Map Reduce model to handle large datasets with accuracy. The experimental result shows that Fisher's exact test gives 98% higher accuracy and also 0.5 sec lesser time than the existing feature selection methods. The selected attributes pH value, Phosphorus, Potassium, Nitrogen, Iron, Sulphur, Boron, and Organic Carbon are considered. Our future work involves an investigation on various types of datasets with the both continuous, discrete values attributes.

ACKNOWLEDGMENT

The First author is thankful to the Kalasalingam Academy of Research and Education Management for providing research fellowship and also sincere gratitude to National Cyber Defence Research Centre (NCDRC) for providing laboratory facilities during this research work.

REFERENCES

1. H. Liu and H. Motoda, Computational methods of feature selection. CRC Press, 2008.
2. I. Polaka, "Feature Selection Approaches in Antibody Display Data Analysis," in 8th International Scientific and Practical Conference, 2011, vol. 2, pp. 16–23.
3. A. e E. Isabelle Guyon, "An Introduction to Variable and Feature Selection," J. ofMachine Learn. Res., vol. 3, 2003, pp. 1157–1182.
4. Das, S., "Filters, wrappers and a boosting-based hybrid for feature selection", in: Proceedings of the Eighteenth International Conference on Machine Learning, 2001.
5. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", Journal of Machine Learning Research, 2003, vol.3, pp. 1157-1182.
6. I. M. El-hasnony, H. M. El Bakry, and A. A. Saleh, "Comparative Study among Data Reduction Techniques over Classification Accuracy", International Journal of Computer Applications, 2015, vol. 122 (2), pp. 8–15.
7. O. Villacampa, "Feature Selection and Classification Methods for Decision Making : A Comparative Analysis," Nova Southeastern University, 2015.
8. R. Porkodi, "Comparison of Filter Based Feature Selection Algorithms : An Overview", International Journal of Innovative Research in Technology & Science, 2014, pp. 108–113.
9. Ruß G., Kruse R, " Feature Selection for Wheat Yield Prediction", In: Bramer M., Ellis R., Petridis M. (eds) Research and Development in Intelligent Systems XXVI. Springer, London, 2010.
10. Surabhi Chouhan, Divakar Singh, and Anju Singh, "An Improved Feature Selection and Classification using Decision Tree for Crop Datasets", International Journal of Computer Applications, vol. 142 (13), 2016, pp. 5-8.
11. R. Parimala, and R. Nallaswamy, "A Study of Spam E-mail classification using Feature Selection package", Global Journal of Computer Science and Technology, vol.11(7), 2011, pp. 45-54.
12. Pinar Yildirim, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease", International Journal of Machine Learning and Computing, Vol. 5(4), 2015, pp. 258-263.
13. Rajeswari, S., and Suthendran, K, "Chi-Square MapReduce Model for Agricultural Data", Journal of Cyber Security and Mobility, 7(1), 2018, pp. 13–24.

14. John, G. H., Kohavi, R., and Pflegger, K, "Irrelevant features and the subset selection problem", In Machine learning: Proceedings of the Eleventh International Conference, 1994, pp. 121-129.
15. M. A. Hall, "Correlation-based Feature Selection for Machine Learning", The University of Waikato, April, 1999.
16. Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. 20(4), 606–626.
17. I. H. W. and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques, " 2nd Edition, ISBN: 0-12-088407-0, Elsevier, 2005.
18. Sutha, K., and Tamilselvi, J. J. (2015). A review of feature selection algorithms for data mining techniques. IJECS, 7(6), 63.
19. Michael J.A. Berry, and Gordon S. Linoff, "Data Mining Techniques For Marketing Sales And Customer Support", Wiley Publication, Second Edition, 2004.
20. Aditi Mahajan, and Anita Ganpati, "Performance Evaluation of Rule Based Classification Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol.3(10), 2014, pp. 3546-3550.
21. Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." Applied Soft Computing 6.2, 2006, pp. 119-138.
22. Rajeswari, S., Suthendran, K., Rajakumar, K., and Arumugam, S. (2016). An Overview of the MapReduce Model. In International Conference on Theoretical Computer Science and Discrete Mathematics, 312–317.
23. Pérez-Ortiz, M., Torres-Jiménez, M., Gutiérrez, P. A., Sánchez-Monedero, J., & Hervás-Martínez, C. (2016). Fisher Score-Based Feature Selection for Ordinal Classification: A Social Survey on Subjective Well-Being. Hybrid Artificial Intelligent Systems, 597–608.
24. Routledge, R. (2008). Fisher's Exact Test. Wiley Encyclopedia of Clinical Trials. doi:10.1002/9780471462422.eoct959
25. P. Somol, P. Pudil and J. Kittler, "Fast branch and bound algorithms for optimal feature selection", IEEE Trans. Pattern Anal. Machine Intell, vol. 26(7), 2004, pp. 900-912.
26. R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
ISBN-3-900051-07-0, URL - http://www.R-project.org/.
27. Karatzoglou , A., Smola, A., Hornik, K,, Zeileis,A.,2005,―kernlab – Kernel Methods.‖ R package, Version 0.6-2. Available from http://cran.R-project.org.
28. Soil health stauts – http://soilhealth.dac.gov.in
29. J. H. and M. Kamber, "DataMining: Concepts and Techniques., " 2nd Edition, Elsevier, 2006.
30. Chouhan, S., Singh, D., and Singh, A. (2016). An Improved Feature Selection and Classification using Decision Tree for Crop Datasets. Int. J. Com. Appl., 142(13), 5–8.
31. Rajeswari S, Suthendran K, and Rajakumar K. (2018). A Smart Agricultural Model by Integrating IoT, Mobile and Cloud-Based Big Data Analytics, International Journal of Pure and Applied Mathematics, 118 (8), pp. 365-370.
32. Rajeswari S, and Suthendran K. (2019). C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud, Computers and Electronics in Agriculture, 156, pp.530-539.

AUTHORS PROFILE

S. Rajeswari received her B.Com (Computer Applications) from Madurai Kamaraj University in 2012; Master of Computer Applications and M.Phil (Computer Science) from Madurai Kamaraj University in 2015 and 2016 respectively. Now, she is a Research Scholar in the Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India. Her current research areas include Big Data, Predictive Analytics, and Data mining.

Suthendran Kannan received his B.E. Electronics and Communication Engineering from Madurai Kamaraj University in 2002; his M.E. Communication Systems from Anna University in 2006 and his Ph.D Electronics and Communication Engineering from Kalasalingam University in 2015. He was a Research and Development Engineer at Matrixview Technologies Private Limited, Chennai for a couple of years. He is now the Head, Cyber Forensics Research Laboratory and Associate Professor in Information Technology, Kalasalingam Academy of Research and Education. His current research interests include Cyber Security, Communication System, Signal Processing, Image Processing, etc.