Chi-Square MapReduce Model for Agricultural Data

S. Rajeswari[1] and K. Suthendran[2]

[1]*Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnan koil - 626126, Tamilnadu, India*
[2]*Department of Information Technology, Kalasalingam Academy of Research and Education, Krishnan koil - 626126, Tamilnadu, India*
E-mail: rajeswari@klu.ac.in; k.suthendran@klu.ac.in

Received 11 January 2018; Accepted 14 March 2018;
Publication 27 March 2018

Abstract

Nowadays, agriculture plays a very significant role in economic growth. Decision making, crop selection and crop yield are the important issues in agriculture productions. Agricultural automation has lead to an incredible growth of software and applications to access the information. Agriculture database contains the farmer's details, land details, soil nutrient details, water levels details and etc. When the data set contains irrelevant, redundant and noisy data then it degrades the performance of the classifier model. The feature selection algorithm is used to improve the performance by selecting the relevant attributes and removing the irrelevant attributes from the database. In this paper, a novel idea is proposed by deploying chi-square technique in MapReduce model to handle large amount of agricultural data. The experimental results show that the proposed Chi-Square MapReduce model has high accuracy and less processing time than the existing feature selection methods.

Keywords: Agriculture, Soil fertility, Attribute selection, Data mining algorithm, filter method and wrapper method.

Journal of Cyber Security, Vol. 7_1, 13–24.
doi: 10.13052/jcsm2245-1439.712

1 Introduction

Data mining is an interdisciplinary field. It is used to extract the hidden information from the data. It have huge amount of data in various fields causes extraordinary challenges for data mining techniques. Researchers and practitioners are realizing that the attribute selection is an integral component for the effective use of data mining tools and techniques [1].

A feature refers to the characteristic of data. Attribute Selection (AS) is used to choose a small subset from the original attributes by following criterion i.e., it is a process of selecting M attributes from the original set of N attributes, $M \subseteq N$. It is one of the essential and indispensable data pre-processing techniques in various domains i.e. artificial intelligence, data mining and machine learning. Attribute significance is classified as strong significance, weak significance and irrelevant [2]. In the data set, strong significance is an optimum subset of attributes which are to be the necessary for the prediction are cannot be removed. The optimal subsets of attributes which are necessary to be chosen based on the certain conditions are known as weak significance. The irrelevant attributes are to be removed because it does not have any information to the target.

A feature which takes the role of another is said to be redundant. Eliminating them reduces the amount of data and lead to improvement in the classification accuracy. It also reduces the dimensionality of feature space and also reduces the running time of the learning algorithm and helps to improve the quality of the classification algorithm [3]. The main idea is used to compare three feature selection methods on agricultural soil data to select the significant Attributes. These attributes are used to improve the accuracy and time taken for the classifier model.

2 Related Work

Feature selection is the only method to find out the important attributes in the data set. It gives the high accuracy rate and gives less time. Here, there are number of authors were carried out several feature selection algorithms. The recent research studies witness that various data mining techniques are used for analysizing agricultural and biological datasets resulted useful classification patterns.

Surabhi Chouhan *et al.* proposed a novel method by using PSO-SVM technique the feature selection from the dataset was done and then using Fuzzy Based Decision Tree classification was also performed. This proposed

approach has been used for Mushroom and Soyabean datasets. The experimental results show that proposed method has the higher accuracy than the existing methods [4].

Ehsan Bijanzadeh [5] has reported that the important features contributing to wheat grain yield is determined by supervised feature selection algorithm. From Iran, 472 fields which are different in 21 characteristics are selected for feature selection process. From the results, the wide range of feature selection created increased the accuracy of the system.

Khan *et al.* utilized the feature selection technique to select an optimal subset from original data set. Using group structure technique they performed the filtering. This improved the accuracy and achieved relatively better classification performance [6].

Hong Wang *et al.* proposed a novel bacterial algorithm for feature selection in classification using control mechanisms and modified population updating strategies. Three parameters have been chosen to limit the randomness of the population and reduced the computational complexity by avoiding the redundant search for the optimal. The proposed algorithm is used to select the best feature subsets on datasets with varying dimensionality and compared with five bacterial based algorithms [7].

The above literature has witness that the accuracy of feature selection methods is an important issue and all the authors have worked for the same. However, still there is a scope for improving the accuracy with reduced processing time. In this work, a novel feature selection model has been proposed to achieve the same.

3 Proposed Methodology

Feature selection is one of the pre processing techniques to remove irrelevant and redundant attributes for the reason of increasing accuracy [8]. It does not imply the cardinality reduction but also the choice of attributes which could be based on presence or lack of interaction among the attributes and the classification algorithm. It is necessary because the high dimensionality and vast amount of data poses a challenge to the learning task.

During the learning process the irrelevant features are to become computationally complex, over fit, become less comprehensible and decrease learning accuracy. Attribute selection methods can be categorized into Filter methods, wrapper methods, embedded methods and hybrid methods. In that, Filter methods can be categorized into univariate and multivariate. Univariate filter methods ignore feature dependencies which can lead to selection of redundant

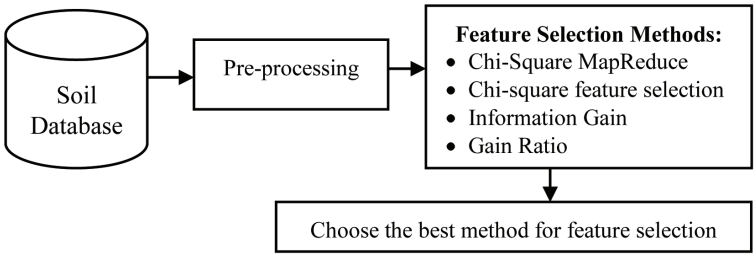

Figure 1 Proposed Method Work flow.

features and worst classification performance when compared to other feature selection techniques [9].

The proposed method workflow is shown in Figure 1. It has the data pre-processing steps to remove the noisy values in the dataset and then select the relevant features for predict the soil fertility level by using attribute selection methods. In that, discover the finest attribute selection method compared with four feature selection methods.

3.1 Data Collection

Data was collected from the http://soilhealth.dac.gov.in during the period 2015–2016 for Virudhunagar district, Tamilnadu. In the above website, it contains the soil testing report taken from 11 blocks of Virudhunagar district. Soil dataset have 15 attributes and a total 25000 instances from the soil sample test report. Table 1 shows the following attributes description for each soil sample test results.

3.2 Data Pre-processing

Raw data contains missing values, noisy data and irrelevant data and unknown data. So, the data pre-processing is the essential step for cleaning the data. During this raw data are transformed into the reasonable format. Nowadays, data causes many errors at the time of analysis because it contains inconsistent, incomplete and noisy values. Using the pre-processing technique, this problem is solved and the data are taken for further processing.

3.3 Feature Selection Method

Feature Selection also called as attribute and variable selection. It is one of the processes to choose a subset of significant attributes. The main goal to discover

Table 1 Soil data set attributes

Attributes	Description
Sample No	Soil Testing Report Identification Number
pH	Soil pH value
EC	Electrical conductivity/mmhos/cm
OC	Organic Carbon
N	Nitrogen/ppm
P	Phosphorus/ppm
K	Potassium/ppm
S	Sulphur/ppm
Zn	Zinc/ppm
Fe	Iron/ppm
Cu	Copper/ppm
Mn	Manganese/ppm
Ca	Calcium/ppm
B	Boron/ppm
Class	Very high, High, Medium, Low, Very Low

the important features to produces the higher performance of classification model [10]. Four Filter methods are used to selecting the relevant attributes and discarding the irrelevant attributes, applied on the preprocessed dataset. In that find the best method for feature selection. The best method was applied on the Big Data MapReduce concepts for calculate the time and accuracy rate.

3.3.1 Chi-Square Feature Selection

Chi-square (x^2) is one of the most popular feature selection methods. This method is used to analysis whether the class label or the target is independent of a particular feature or not. This is used to select the predictor variable [11]. This value for attribute with 'r' different values and 'c' number of classes is defined as

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

Where O_{ij} is the number of occurrences with value of 'I' and which are in class 'j'. E_{ij} is the expected number of occurrences with value 'I' and class 'j'.

3.3.2 Information Gain Feature Selection

It is one of the standard methods for feature selection. The purpose of these techniques is to discard irrelevant or Information gain feature selection, entropy value has been calculated for whole data [12]. It is a supervised, univariate, simple, powerful, symmetrical and entropy-based feature selection

algorithm. The Information Gain for a feature X and the class label Y is calculated using the formula.

$$InformationGain(X, Y) = H(X) - H(X|Y) \tag{2}$$

Where H(X), H (X|Y) are entropy values calculated on X and Y. The entropy of X can be calculated as

$$H(X) = -\sum_{i} P(x_i) \log_2(P(x_i)) \tag{3}$$

The entropy of X|Y can be computed as

$$H(X|Y) = -\sum_{j} P(y_j) \sum_{i} P(x_i \mid y_j) \log_2(P(x_i \mid y_j)) \tag{4}$$

This method calculates the ratio on behalf of each and every attributes separately and selects 'm' features as more relevant ones among 'n' features with high Information Gain i.e., it considers a feature F with high Information Gain as the more relevant one. The main drawback of the algorithm is that it selects the feature with high Information Gain which may or may not be more informative. The Information Gain cannot handle redundant features [13] because the features are chosen in a univariate way.

3.3.3 Gain Ratio Feature Selection

Gain Ratio have the subset have been selected using the entropy value as well as the information gain value. From the given equation Gain Ratio subset selector have been produced. It is supervised, univariate, non-symmetrical and entropy based measure introduced to eliminate the bias of Information gain [14]. Gain Ratio can be computed as

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

In order to predict class label Y, it needs to be normalized by dividing Information gain by entropy of feature X and vice versa. Gain Ratio weight value range between the 0 and 1. If value 1 specifies that the knowledge of Y completely predicted by X and 0 specifies that Y and X are uncorrelated. This method favours the features with less value.

3.3.4 Chi-square MapReduce Feature Selection

Chi-Square feature selections have two parameters observed and expected frequency. It was calculated by the MapReduce techniques [15]. Further, it

can be discovered the attributes weights. The highest weight attributes are the relevant attributes.

Pseudo code for Chi-Square MapReduce feature selection
Input: Pre-processed Data
Output: Relevant Attributes

Map Phase:
Step 1: Choose the node attributes in the data set
Step 2: Compute expected frequency for each attributes
Step 3: For each attribute in sequence, Calculate Observed Frequency

Reduce Phase:
Step 1: Sum the observed and expected square values and divided by expected frequency
Step 2: Select attribute with the highest weight to be the next node
Step 3: Remove the node attribute

Combine Phase:
Repeat Map and Reduce phase steps until all attributes have been used.

The above feature selection methods are compared to find the best method for soil data. This method is used to improve the accuracy performance of the classification algorithm.

4 Result and Discussion

R language is an open source software Package and it is mostly used to process the statistical data. It does not handle the large amount of data. So the Rhadoop, RmR, Rhdfs packages are used to integrate the R and Hadoop environment. It works with the terra bytes of data and easily handles the data. The proposed method code was written with the help of the Rhadoop packages. The main objective of proposed model is used to choose the important features for predict soil fertility level with high accuracy and less time. This work is compared with the other feature selection algorithm and the best one is taken into for MapReduce method.

Table 2 show that data set was stored in the CSV (Comma Separated Values) file format. It was collected over the blocks of Virudhunagar District. Soil data set contains 15 attribute. In that, Class is the dependent variable and all other variables are the predictor variables which are used to predict the soil fertility level.

Table 2 Sample soil data set

S. No	Sample No	pH	EC	OC	..	Zn	Fertlilty Level
1	TN641648/2016-17/3342212	8.2	0.1	0.11	..	0.08	Very high fertile
2	TN641648/2016-17/3342269	8.2	0.1	0.05	..	0.13	Very high fertile
3	TN641648/2016-17/3342322	8.2	0.1	0.11	..	0.42	Low fertile
⋮	⋮	⋮	⋮	⋮	..	⋮	⋮
n	TN641653/2016-17/3533704	8.1	0.09	0.19	..	0.2	Very high fertile

In that, preprocessing step result of soil data set. During the preprocessing, the missing values in the soil data set are identified, and removed. Before removing missing values, the raw data contains 25000 records. In that, inconsistent data, error data, and missing values were removed. After preprocessing the soil test report, dataset has 15 attributes and a total of 24980 instances of Virudhunagar district. So, the preprocessed soil data set was taken for further classification process.

Figure 2. shows the weights for the feature selection methods are compared. In that, chi-square MapReduce method gives the relevant attributes compare with the Information gain, chi-square and gain ratio features selection methods. The highest weight attributes are taken into the classification algorithms. The highest weight attributes are pH, K, N, Fe, S, B, and OC taken into the classification algorithms.

Table 3, shows the experimental outcome of four feature selection methods. In that, Chi-square MapReduce gives high accuracy (98%) and also less time (0.5 sec) was taken for find the relevant features was compared with the other existing three feature selection methods.

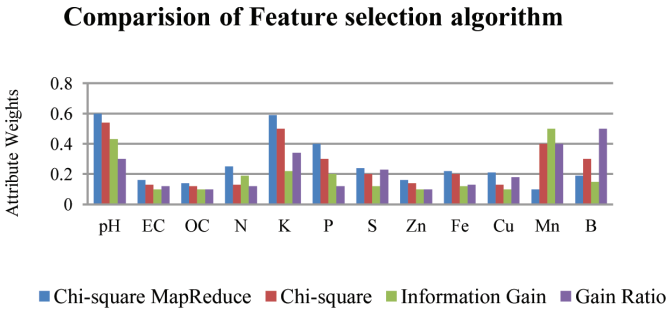

Figure 2 Comparision of Feature selection result.

Table 3 Comparision results of feature selection methods

Feature Selection Methods	Correctly Identified Features	Incorrectly Identified Features	Accuracy (%)
Chi-square MapReduce	10	2	98%
Chi-square feature selection	9	3	97%
Information Gain feature selection	7	5	90%
Gain Ratio feature selection	8	4	91%

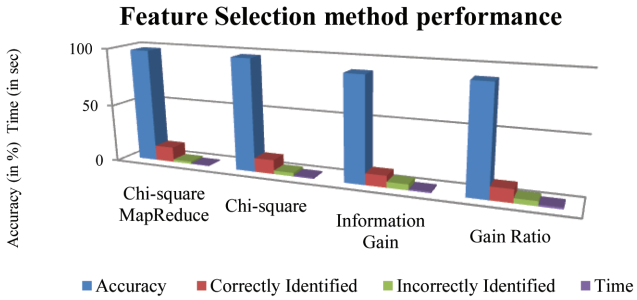

Figure 3 Feature Selection method performance.

Figure 3 shows the comparison result of accuracy, execution time, and correctly, incorrectly identified features of proposed approach with the existing feature selection methods. In that, Chi-square MapReduce produces the relevant features from large data set to predict the soil fertility level.

5 Conclusion

Feature selection has been a most significant research issue due to the availability huge amount of data with thousands of features and it can be widely used in many domains such as statistics, text mining, web mining, machine learning, microarray data analysis and image processing. To meet the objective, a novel method has been proposed with the help of feature selection technique to improve the performance of the classifier by minimizing redundancy, removing noisy data and maximizing the relevance. It also helps us to analyze the methodology behind each algorithm in selecting the more relevant features and removing irrelevant features. In this work, using Chi-square MapReduce model the relevant features are selected quickly comparing with the other feature selection methods. The experimental result shows that chi-square model gives 98% higher accuracy and also 0.5 sec lesser time

than the existing feature selection method. The selected attributes pH, K, N, Fe, S, B, and OC are taken into further classification algorithms. This work may be extended to design a feature selection algorithm for high dimensional multiclass dataset with considerably improvement in accuracy with less space and time requirement in future.

Acknowledgment

The first author is thankful to the management of Kalasalingam University for providing fellowship and also thanks National Cyber Defence Research Centre for supporting laboratory facilities during this research work.

References

[1] Li, Z., Shang, Z., Qu, B. Y., and Liang, J. J. (2014). Feature selection based on manifold-learning with dynamic constraint handling differential evolution. In *Evolutionary Computation (CEC),* 332–337.

[2] Vanaja, S., and Kumar, K. R. (2014). Analysis of feature selection algorithms on classification: a survey. *Int. J. Com. Appl.*, 96(17).

[3] 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.

[4] Kumar, V., and Minz, S. (2014). Feature selection. *SmartCR*, 4(3), 211–229.

[5] Raorane, A. A., and Kulkarni, R. V. (2012). Data Mining: An effective tool for yield estimation in the agricultural sector. *IJETTCS*, 1(2), 75–79.

[6] 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.

[7] Bijanzadeh, E., Emam, Y., and Ebrahimie, E. (2010). Determining the most important features contributing to wheat grain yield using supervised feature selection model. *Australian Journal of Crop Science*, 4(6), 402–407.

[8] Wang, J., Zhao, Z. Q., Hu, X., Cheung, Y. M., Wang, M., and Wu, X. (2013). Online Group Feature Selection. In Proceeding *IJCAI* 1757–1763.

[9] Khan, R. A,, and Mandwi, I. (2017). "A Survey on Multi-Objective Unsupervised Feature Selection Using Genetic Algorithm", *IJIRCCE*, 5(1), 103–108.

[10] Wang, H., and Niu, B. (2017). A novel bacterial algorithm with randomness control for feature selection in classification. *Neurocomputing*, *228*, 176–186.
[11] Sutha, K., and Tamilselvi, J. J. (2015). A review of feature selection algorithms for data mining techniques. *IJECS*, 7(6), 63.
[12] Pino, A., and Morell, C. (2013). Analytical and Experimental Study of Filter Feature Selection Algorithms for High-dimensional Datasets. In *Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support*. Atlantis Press.
[13] Goswami, S., and Chakrabarti, A. (2014). Feature selection: A practitioner view. *International Journal of Information Technology and Computer Science (IJITCS)*, 6(11), 66–67.
[14] Rajeswari, S., Suthendran, K., and Rajakumar, K. (2016). "A Smart Agricultural Model by Integrating IoT, Mobile and Cloud-Based Big Data Analytics", *IEEE Sponsored International Conference on Engineering and Technology*, 4, 82–86.
[15] 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.

Biographies

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.