ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data

R. Revathy, R. Lawrance

Research Scholar, Department of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

Director, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India

ABSTRACT: Data mining is quite finding the hidden information and correlation between the massive data set that is helpful in decision making. Decision tree is one of the predictive modeling approach for representing data that can be in visualizing manner. In the agricultural field, data mining can help farmers to develop yield and country development. By applying data mining techniques, crops can be protected from pests by predicting and enhancing crop cultivation. This paper involves data pre-processing to eliminate noisy data in crop pest data that offers better accuracy. Feature selection takes an essential pre-processing step is to improve the mining performance by reducing the number of attributes. In this paper oneR feature selection is used for filtering crop pest dataset attributes instead of using full attribute set. It finds weights of discrete attributes and treats all numerically valued features as continuous and uses a simple method to separate the range of values into several disjoint intervals. According to the weights, splitting attributes have been chosen for generating decision tree. This paper focuses on the comparison of C4.5 and C5.0 decision tree algorithms for pest data analysis with an experimental approach. C5.0 proved its efficiency by giving more accurate result rapidly and holding less memory while comparing c4.5 algorithm.

KEYWORDS: Data Mining, Data preprocessing, OneR feature selection method, C4.5 Classification, C5.0 Classification.

I. INTRODUCTION

Data mining is the process of using large data sets to gather important hidden knowledge. It is divided into seven steps like data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining is used for any kind of data, including database data, data warehouse data or transactional data [3].

Agriculture sector plays a crucial role in our economy. More than 70 percentage of the population depends on agriculture or agriculture practice. The agricultural yield is mainly depends on weather conditions, diseases and pests, planning of the harvest operation [1]. Due to contrast of climate factors the agricultural productivities in India are continuously decreasing over a decade. Data mining in agriculture is a novel research field that allows predicting the future trends and behaviors. Farmers are not only harvesting crops, but also harvesting large amount of raw data. Data mining provides the methodology to transform these raw data into useful information for decision making to get better crop management [2]. Data Mining has been used to analyze large datasets and establish useful classification of crop pest data. This paper comprises two different classification techniques and approaches in order to predict the crop pests.

II. LITERATURE REVIEW

Literature survey obtains an understanding of the fundamentals and learning the definitions of the concepts. The intend of the literature survey is accessing latest approaches, methods, theories and discovering a new research based on the existing research.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Novakovic, J., [4] discusses that OneR method evaluates features individually and ranks according to error rate (on the training set). It treats all numerically valued features as continuous and produces simple rules based on one feature only. It handles missing values by treating "missing" as a legitimate value.

Kaur, D., *et al.*, [6] evaluate that C4.5 algorithm is enhanced to ID3. C4.5 can handle continuous input attribute. It makes a Splitting of categorical attribute which is similar to ID3 algorithm. Continuous attributes always generate binary splits and attribute with highest gain ratio is selected. It avoids over fitting of decision tree by providing the facility of pre and post pruning.

Patil, N., *et al.*, [5] state that C5.0 decision tree is better than C4.5 decision tree on the efficiency and the memory. C5.0 decision tree works by splitting the sample based on the field that provides the maximum information gain. It can split samples on the basis of the biggest information gain field.

Revathi, P., *et al.*, [7] refer that decision tree widely used learning method and do not require any prior knowledge of data distribution, works well on noisy data .It has been applied to classify Rice disease based on the symptoms. It discovered classification rules for the Indian rice diseases using the c4.5 decision tree algorithm.

Lawrance, R., *et al.*, [16] examine the predictive mean matching method for imputing the missing values for continuous variables. It imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value for the missing value.

This literature review provides a flexible study about data mining preprocessing techniques, feature selection method, different classification and prediction. Based on this survey, this research embraces of three parts.

The first part includes preprocessing like,

- Elimination of dirty data in crop pest data set

The second part includes Feature Selection such as,

- OneR method

Third part includes comparison of two different Classification like,

- C4.5 decision tree
- C5.0 decision tree

III. DATA PREPROCESSING

Data preprocessing has been often neglected but important and a prerequisite step in the data mining process. Low quality data will lead to low quality mining results. Thus the data can be preprocessed in order to improve the quality of the data. Data quality can be accessed in terms of accuracy, completeness and consistency.

Preprocessing reconstructs the data into a format that will be very easy and effective for further processing [12]. There are various tools and techniques that are used for preprocessing which includes:

Data cleaning: It can be applied to remove noise and correct inconsistencies in data.
Data integration: It merges data from multiple sources into a coherent data store such as a data warehouse.
Data reduction: It can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.
Data transformation: It may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0 [11].

Unwanted data can cause confusion for the mining procedure, resulting in unreliable output. In this paper some of the useless crop pest data have been eliminated by the data filtered method. This can improve the accuracy and efficiency of the decision tree.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Avg_SpoEggMass	Avg_SpoGregLar	Avg_SpoSolLar	Avg_semilooper	Avg_harmigera
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.6	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.6	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.6	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0

Fig. 1. Before preprocessing

Fig. 1. represents the crop pest dataset that includes noisy data.

Avg_SpoEggMass	Avg_SpoGregLar	Avg_SpoSolLar	Avg_semilooper	Avg_harmigera
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.6	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.6	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.2	0.0	0.0	0.0	0.0
0.6	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.4	0.0	0.0	0.0	0.0
0.0	0.0	0.2	0.0	0.0
0.0	0.0	0.2	0.0	0.0

Fig. 2. After preprocessing

Fig. 2. represents the removal of noisy data from crop pest dataset.

IV. FEATURE SELECTION

Feature selection is a technique for identifying a subset of features by removing irrelevant or redundant features. The importance of feature selection is reducing the cost of learning by reducing the number of attributes.It provides better learning performance compared to using full attribute set.

Before applying the algorithm, irrelevant attributes need to be filtered so as to get better the efficiency of the classification algorithms. There are two approaches for attribute selection, namely the filter approach and the wrapper approach The filter approach uses measures such as attributes weights, consistency or distance measures to compute the relevance of a set of features while the wrapper approach uses the predictive accuracy of a classifier as a means to evaluate the “goodness” of a feature set.

Feature selection techniques provide three main benefits when constructing predictive models:

- Improved model interpretability
- Shorter training times

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

- Enhanced generalization by reducing over fitting [10].

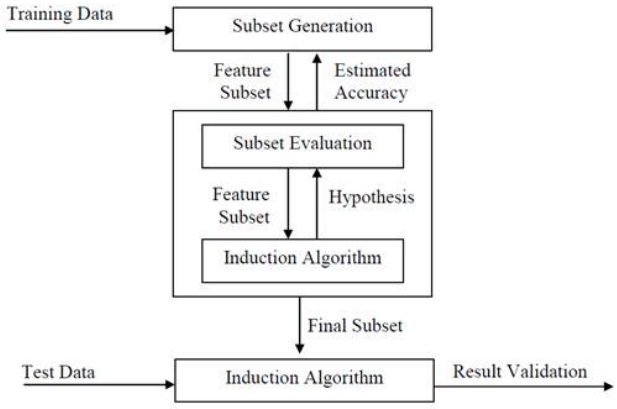

Fig. 3. Working of feature selection

The complete feature set can be evaluated by filter feature selector to generate feature subset as shown in Fig. 3. The feature subset can further evaluate by hypothesis to get classification model.

This paper proposes OneR feature selection method for attribute selection by calculating weights.

OneR Method

OneR algorithm finds weights of continuous and discrete attributes. It was first described by Holte in 1993 which is simple, fast, and cheap method to attribute weighing. By using this method, the attributes of the crop pest data are evaluated the features individually.

It ranks the features according to error rate (on the training set). It makes rules that test a single attribute and branch accordingly (each branch corresponds to a different value for that attribute). It handles missing values by treating "missing" as a legitimate value.This is one of the most primitive schemes which produces simple rules based on one feature only. Although it is a minimal form of classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes [4].

Pseudo code for OneR Algorithm

```
for each attribute
{ for each value of that attribute
{ compute the class distribution based on attribute value
Class_label = select most frequent class
create a rule: attribute = value => Class_label
}
Calculate the error rate of the rule on the whole dataset
}
Select rule with lowest error rate
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

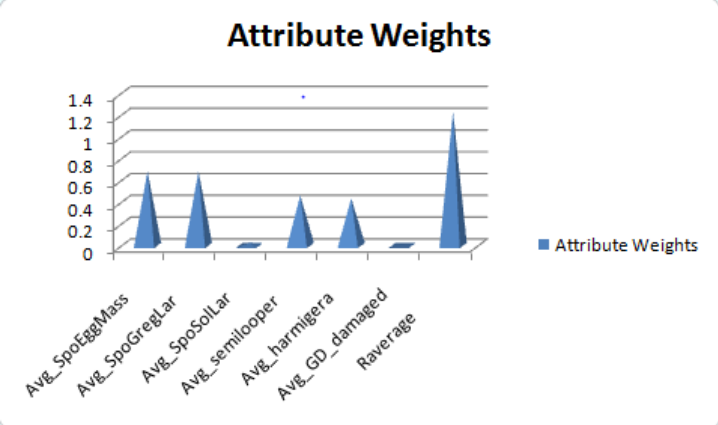

Fig. 4. Weights derived by OneR method

The weights of the different attributes of crop pest data are displayed in the above Fig. 4.

Although each feature method has its own unique features, OneRalgorithm is efficient and can correctly estimate the quality of attributes with strong dependencies between attributes and eliminate the redundant features.This paper carried out the experiments on crop pest data with OneR method to select important features which improve the accuracy of the classification. Here, Raverage attribute holds a maximum weight compared to all other attributes.

V. CLASSIFICATION

Classification is a task that occurs very frequently, which involves dividing up objects so that each is assigned to one a number of exhaustive and exclusive categories known as classes. Many practical decision making tasks can be formulated as classification problems.

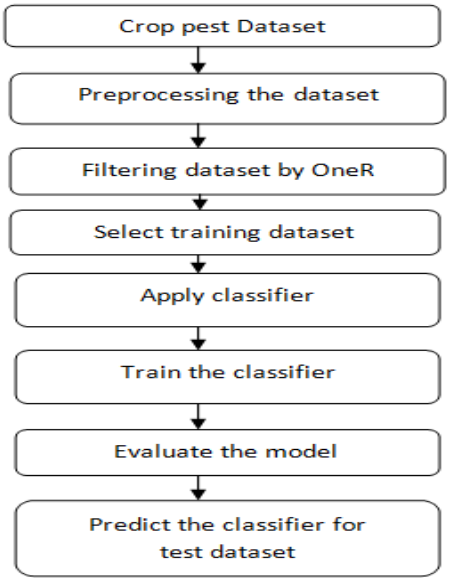

Fig. 5. Framework for classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

The flow of the phases in classification is displayed in the above Fig. 5. It represents how the training data are preprocessed and evaluate the model

This paper has split the crop pest dataset into training and testing data set. A model is built from the training data set which value of class label is known. Classification algorithms are used to create models from training data sets. Accuracy of the model is predicted by test data set. Among classification algorithm, decision tree algorithms are usually used because it is easy to follow and economical to implement [6].

Decision Trees

The decision tree is quite common modelling method to classify, since it is a classification technique. The decision tree is comprised of nodes that form a rooted tree that means a directed tree with a node called "root" that has no incoming edges. All other nodes have precisely one incoming edge. A node with outgoing edges is known as internal or test node. All other nodes are termed as leaves (also known as terminal or decision nodes) [13].

In this paper, the decision tree is built from the training data set and the testing data set is used to predict the accuracy of the decision tree. Here, the attribute with maximum weight is selected as splitting attribute.

A. C4.5 CLASSIFICATION

C4.5 is an extension of ID3, developed by Quinlan in 1993. This algorithm produces a decision tree for the given crop pest training data by recursively splitting that data. The decision trees generated by C4.5 can be used for classification since it is often referred to as a statistical classifier [8].

Algorithm to generate C4.5 decision tree

Input: an attribute –valued dataset *D*

1. Tree={}
2. if *D* is "pure" or stopping criteria met then
3. terminate
4. else if
5. for all attribute $a \epsilon D$ do
6. Compute information-theoretic criteria if we split on *a*
7. end for
8. a_{best} = Best attribute according to above computed criteria
9. Tree = Create a decision node that tests a_{best} in the root
10. D_v = induced sub –datasets from *D* based on a_{best}
11. For all D_v do
12. $Tree_v$= C4.5(D_v)
13. Attach $Tree_v$, to the corresponding branch of tree
14. end for
15. return Tree

Algorithm : Generate C5.0 decision tree

C4.5 decision tree grows using Depth-first strategy. It allows pruning of the resulting decision trees. It can also deal with numeric attributes, missing values, and noisy data. In order to handle continuous attributes, it creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

B. C5.0 CLASSIFICATION

C5.0 algorithm is an extension of C4.5 algorithm which is also an extension of ID3. It is the classification algorithm which is suitable for big data set. It is improved than C4.5 on the speed, memory and the efficiency. C5.0 model works by splitting the crop pest training data that provide the maximum weight. C5.0 is easily handled the multivalue attribute and missing attribute from crop pest training data set. In this paper the training pest data is used for constructing C5.0 decision tree while the testing pest data is used for prediction [8][9].

Algorithm to generate C5.0 decision tree

Input

a. Data partition, *D*, a set of training tuples and their associated class labels
b. attribute_list, the set of candidate attributes
c. *attribute_selection_method*, a procedure to determine the splitting criterion partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and ,either a *split-point* or *splitting subset*

ouput: C5.0 decision tree

Method:

1. create a node N
2. if tuples in *D* are all of the same class, *C*, then
3. return *N* as a leaf node labelled with the class *C*
4. if *attribute_list* is empty, then
5. return *N* as a leaf node labelled with the majority class in *D*
6. apply *attribute_selection_method(D, attribute_list)* to find the best *splitting_criterion*
7. label node *N* with *splitting_criterion*
8. if *splitting_attribute is discrete-valued and multiway splits allowed then*
9. *attribute_list* ←*attribute_list- splitting_attribute*
10. For each outcome *j* of *splitting_criterion*
    Let D_j be the set of data tuples in *D* satisfying outcome *j*
    if D_j is empty then attach a leaf labelled with majority class in *D* to node *N*
    else, attach the node returned by Generate C5.0 decision tree(D_j, *attribute_list*) to node *N*
11. Return *N*

In this paper, C4.5 classification and C5.0 classification have been implemented on the crop pest training dataset and compared with its accuracy. Even though C5.0 is similar to C4.5, C5.0 handles all types of data like continuous,categorical, dates, times and timestamps. It can deal with missing values of crop pest data. It mainly supported boosting to improve the classifier accuracy.

VI. RESULTS AND DISCUSSION

This research is performed on the agricultural crop pest data set which is composed of 16 attributes and 1000 instances.

R tool is used here for implementation of C4.5 algorithm and C5.0 algorithm. R is a statistical programming language which is a free open source package based on the S language developed by Bell Labs. The language is very powerful for writing programs. Many statistical functions are already built in. Corresponding contributed packages expand the functionality to cutting edge research.

C4.5 and C5.0 decision tree has been constructed according to pest training data and compared according to prediction of testing dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

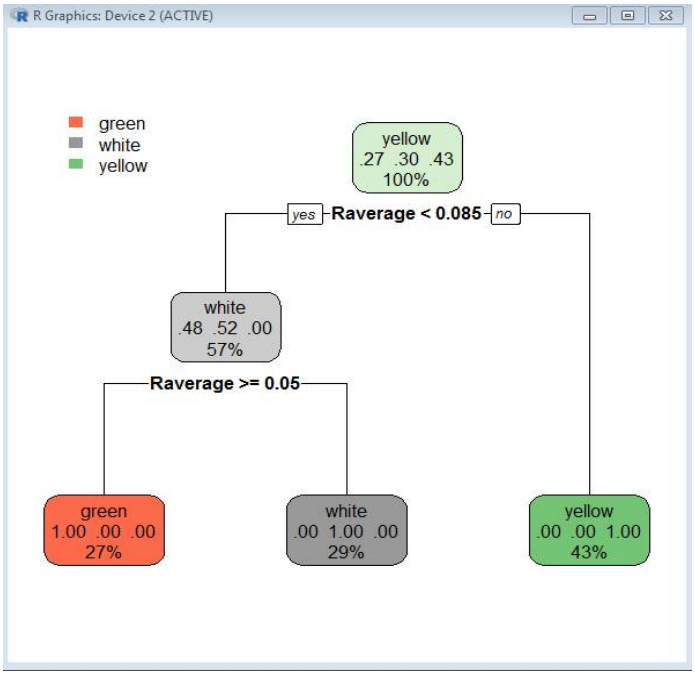

Fig. 6. Decision tree for C4.5 Algorithm

Fig. 6. represents the C4.5 decision tree for the crop pest training data. This tree signifies the classification of pest which is categorized by color.

In C4.5 model, 98% of the data are correctly classified. The accuracy rate is predicted by a test data set which is up to 98.48%. It obtains the error rate of 1.52%.

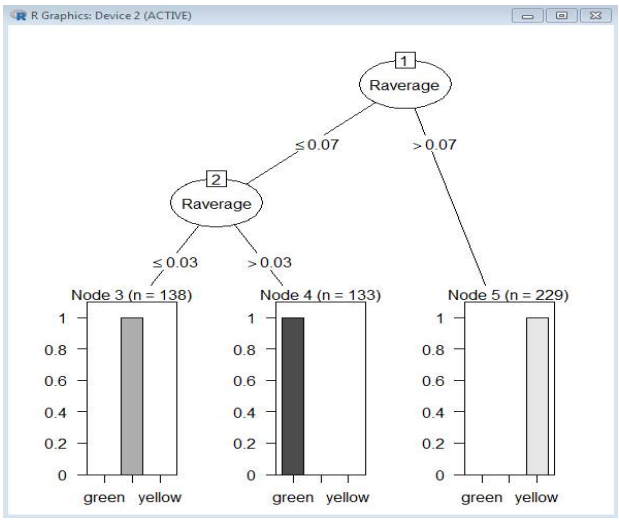

Fig. 7. Decision tree for C5.0 Algorithm

Fig. 7. represents the C5.0 decision tree for the crop pest training data and classifies the pest by color. 99.49% of data are correctly classified in C5.0 model. The error rate in C5.0 is measured as 0.51%.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 1, March 2017

Classifier	*C4.5*	*C5.0*
Correctly classified instances	195	197
Incorrectly classified instances	3	1
Accuracy Prediction	98.48%	99.49%
Total time taken to build model (in seconds)	0.02	0.01
Error Rate	1.52%	0.51%

Table 1. Experimental Results

Table 1. embraces the result of the decision tree, i.e., correctly classified instances, incorrectly classified instances, accuracy, error rate and duration of the decision tree.

This research proved the efficiency of the C5.0 algorithm since it predicted more accuracy, short duration and less error rate as compared to the C4.5 algorithm

VII. CONCLUSION

This paper compared C4.5 and C5.0 decision tree algorithms with an experimental approach. It illustrated the proficiency of C5.0 classifier with experimental results as shown in the Table 1, which confirmed that the most powerful and preferred method in machine learning is obviously C5.0. As the size of dataset becomes extremely big, the process of building a decision tree can be quite time consuming and data cannot fit in memory. To overcome this difficulty, C5.0 with Map Reduce algorithm will be executed in the future.

REFERENCES

[1] *Raorane, A, A., Kulkarni, R, V.,* "Data Mining: An effective tool for yield estimation in the agricultural sector", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume.1, Issue.2, August 2012, pp: 75-79.
[2] *Nasira, G, M., Hemageetha, N.,* "Vegetable Price Prediction Using Data Mining Classification Technique", Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 2012, pp: 100-102.
[3] *Mehta, S, T., Kathiriya, R, D.,* "Survey of Data Mining Techniques in Precision Agriculture", IJSR - INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH, Volume.4, Issue72, July 2015, pp: 363-364
[4] *Novakovic, J.*, "The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier", 18th Telecommunications forum TELFOR 2010, November 2010, pp: 1113-1116.
[5] *Patil, N., Lathi, R.,Chitre, V.,* "Comparison of C5.0 & CART Classification algorithms using pruning Technique", International Journal of Engineering Research & Technology (IJERT), Volume.1, Issue.4, June 2012, pp: 1-5.
[6] *Kaur, D., and Bedi, R.,* Gupta, K, S., "Review of Decision Tree Data Mining Algorithms: ID3 AND C4.5", International Conference on Information Technology and Computer Science, July 2015.
[7] *Revathi, P., Revathi, R., Hemalatha, M.,* "Comparative Study of Knowledge in Crop Diseases Using Machine Learning Techniques, Volume. 2, Issue.5, June 2011, pp: 2180-2182.
[8] *Krishna Kumar, V, S., Kiruthika, P.,* "An Overview of Classification Algorithm in Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Volume. 4, Issue.12, December 2015, pp: 255-257.
[9] *Patel, R, B., Rana, K, K.,* "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research (IJEDR), Volume. 2, Issue.1, 2014, pp: 1-5.
[10] *Pundir, L, S., and Amrita.,* "FEATURE SELECTION USING RANDOM FOREST IN INTRUSION DETECTION SYSTEM", International Journal of Advances in Engineering & Technology(IJAET), Volume. 6, Issue.6, July 2013, pp: 1319-1324.
[11] *Hen J. and Kamber M.,* "Data Mining: Concepts and Techniques, Second Edition, ELSEVIER Publications, ISBN: 978-81-312-0535-81, 2005.
[12] *Bharat, V., Shelale, B., Khandelwal, K., Navsare, S.,* "A Review Paper on Data Mining Techniques", International Journal of Engineering Science and Computing (IJESC), Volume. 6, Issue.5, May 2016, pp: 6268-6271..
[13] *Singh, S., and Gupta, P.,* "COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY", International Journal of Advanced Information Science and Technology (IJAIST), Volume. 27, Issue.27, July 2014, pp: 97-103.
[14] *HSSINA, B., MERBOUHA, A., EZZIKOURI, H., and ERRITALI, M.,* "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 2014, pp: 13-19.
[15] *Devi, C, J.,* "Binary Decision Tree Classification based on C4.5 and KNN Algorithm for Banking Application", International Journal of Computational Intelligence and Informatics, Volume. 4, Issue.2, September 2014, pp: 125-131.
[16] *Jinubala, V., and Lawrance, R.,* "Analysis of Missing Data and Imputation on Agriculture Data With Predictive Mean Matching Method", International Journal of Science and Applied Information Technology (IJSAIT),, Volume.5, Issue.1, 2016, pp: 01-04.
[17] *Sutha, S., Tamilselvi, J, J.,* "A Review of Feature Selection Algorithms for Data Mining Techniques", International Journal on Computer Science and Engineering (IJCSE), Volume. 7, Issue.6, June 2015, pp: 62-67.