World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:10, No:5, 2016

# Performance Comparison of ADTree and Naive Bayes Algorithms for Spam Filtering

Thanh Nguyen, Andrei Doncescu, Pierre Siegel



***Abstract*—**Classification is an important data mining technique and could be used as data filtering in artificial intelligence. The broad application of classification for all kind of data leads to be used in nearly every field of our modern life. Classification helps us to put together different items according to the feature items decided as interesting and useful. In this paper, we compare two classification methods Naïve Bayes and ADTree use to detect spam e-mail. This choice is motivated by the fact that Naive Bayes algorithm is based on probability calculus while ADTree algorithm is based on decision tree. The parameter settings of the above classifiers use the maximization of true positive rate and minimization of false positive rate. The experiment results present classification accuracy and cost analysis in view of optimal classifier choice for Spam Detection. It is point out the number of attributes to obtain a tradeoff between number of them and the classification accuracy.

***Keywords*—**Classification, data mining, spam filtering, naive Bayes, decision tree.

## I. Introduction

DATA Mining allowed the development of a new research field "The Big Data". The term "Big Data" is the successor of "information explosion" term. The "Big Data" was appeared for the first time by John Mashey in 1998 [1]. "Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze [2]. This new field tries to answer how a huge number of databases and information repositories could be organized, analyzed and how it is possible to retrieve information from this data. It is obviously these questions generate an eminent need of methods that can help users to efficiently navigate, summarize, and organize the data so that it can further be used for applications ranging from market analysis, fraud detection [3].

The Internet development involves the new technics of data storage on distant server called clouds. The emails are used so that the total email traffic worldwide, including emails professionals and individuals was estimated at over 144 billion emails per day at the end of the year 2012. It is also expected that the amount of mail traffic reaches more than 192 billion e-mails a day in 2016 [4]. Some of these e-mails are promotions and could be considered as not interesting therefore as SPAMS.

In this paper, we analyze some known data results may uncover important data patterns are needed.

Thanh Nguyen and Andrei Doncescu are with the University of Toulouse, Toulouse, France (e-mail: tnguyen@laas.fr, andrei.doncescu@laas.fr).

Pierre Siegel is with the LIF-AIX Marseille University, Marseille, France (e-mail: pierre.siegel@cim.univ-mrs.fr).

## II. Data Mining

Data mining is an analytical process designed for extracting or exploring hidden and predictive information from large databases. It can also be described as the process of searching for valuable information in large volumes of data [5].

Data mining is a form of knowledge discovery essential for solving problems in a specific domain, means a process of nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [6].

Data mining is widely used in diverse areas like Financial Data Analysis, Telecommunication Industry, Biological Data Analysis, Intrusion Detection and other Scientific Applications. Data mining refers to the analysis and extracts knowledge from the large quantities of data that are stored in computers, network and internet [3].

Data mining should be applicable to any kind of information repository from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, hypertext documents, relational databases, object-relational databases, object oriented databases, data warehouses, transaction databases, unstructured and semi-structured repositories such as the World Wide Web, multimedia databases, time-series databases etc. [7]. These functions of data mining are mainly classified as include clustering, classification, prediction, associations and sequential patterns [8].

In this paper, we focus research on the Spam data classification and the performance measure of the two classifier algorithms ADTree and Naive Bayes based on True Position Rate (TP Rate), False Position Rate (FP Rate) generated by the algorithms when applied on the Spambase data set.

## III. Spam Classifiers

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome called prediction attribute. The algorithm analyses relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute not yet known. The algorithm analyses the input and produces a prediction [9]

In this section, it is presented two types of algorithm: Naive bayes classifiers algorithm and ADTree decision tree algorithm in the view of comparison. The comparison is made on accuracy, sensitivity and specificity using true positive and

false positive in confusion matrix generated by the respective algorithms. As well as, correct and incorrect instances that helping to define a most efficient classification method by using the confusion matrix.

### *A. Naive Bayes Algorithm*

Naive Bayes algorithm belongs to text retrieval methods and intends to assign class labels to problem instances using a traditional probabilistic model. The success of this technic is due to the text retrieval application, which it explains the natural application to spam filtering. The choice of this classifier is suited when the dimensionality of the input is high and it requires a small amount of training data to estimate the intrinsic parameters.

The Naive Bayes algorithm calculates a set of probabilities, in a very traditional from normalized histogram. It assumes that all attributes are conditional independents and the probability of a vector x belongs to a class C is calculated by:

$$p(C_k|x_1, \dots, x_n) \tag{1}$$

This means that under the above independence assumptions, the conditional distribution over the class variable $C$ is:

$$p(C_k|x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{2}$$

where the evidence $Z = p(x)$ is a scaling factor dependent only on $x_1, \dots, x_n$, that is, a constant if the values of the feature variables are known. Even, if this independence assumption is rarely true for "real data" the algorithm tends to perform well and learn rapidly in various supervised classification problems [10].

### *B. ADTree Algorithm*

An Alternating Decision Tree (ADTree) [11], [12] is a generalization of decision trees and voted-stumps. As well as decision tree an ADTree consists of an alternation of binary decision nodes and prediction leaves. Each leaf is a logical formula (CNF) of the pathway node decisions with weights.

An instance is classified by an ADTree algorithm by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed. An instance is scored by summing the scores of all decision. This is different from binary classification trees such as CART (Classification and regression tree) or C4.5 in which an instance follows only one path through the tree [13].

The ADTree algorithm produces a set $P = P_i$ of preconditions and a set $R = R_i$ of rules. A single rule consists of a simple conditional involving a precondition, a test condition $c_i$ and a set of signed numerical predictions $p_1$ and $p_2$. The set of all test conditions is labeled $C$. Preconditions are conjunctions of conditions and negations of conditions.

Input: Precondition $P \in \mathcal{P}$, Condition $c_1 \in \mathcal{C}$, Scalars $p_1, p_2 \in \mathbb{R}$
Result: Number (either $p_1$ or $p_2$ (or 0) denoting a single prediction)
if $P$ then
    if $c_1$ then
        return $p_1$
    else
        return $p_2$
    end
else
    return 0
end

The ADTree algorithm consists of an alternation of decision nodes, which specify a predicate condition, and prediction nodes. ADTrees always have prediction nodes as both root and leaves. An instance is classified by following all paths for which all decision nodes are true, and summing any prediction nodes that are traversed.

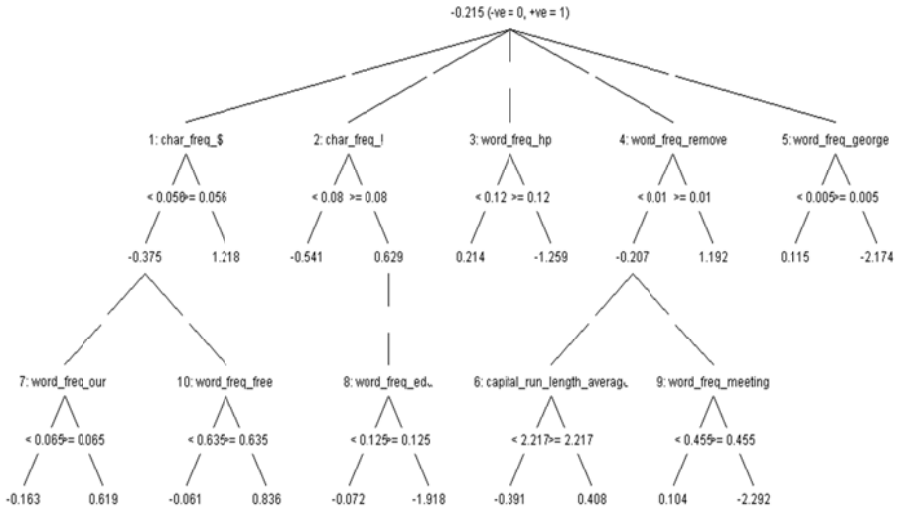


Fig. 1 Constructed tree using ADTree on the Spambase dataset

## IV. Performance Evaluation and Comparison

The function which compares classification models by providing a quality measure for classifier when solving a classification problem is score. The score is based on Confusion matrix corroborates with Receiver Operating Characteristics (ROC).

The measurements obtained by using Confusion Matrix are: Accuracy, recall, specificity, precision and F-score and in the case of ROC I the area under the ROC curve (AUC). Independently of the indices adopted, an important aspect to be considered is the asymmetry in the misclassification costs.

A Spam message incorrectly classified as legitimate is a relatively minor problem, as the user is simply required to remove it. On the other hand, a legitimate message mislabeled as Spam can be unacceptable, as it implies the loss of potentially important information, particularly in configurations in which Spam messages are automatically deleted. For this reason, describing the performance of an algorithm solely in terms of the classification accuracy (the relative number of messages correctly classified) is not adequate, as it assumes equal misclassification costs for both classes.

We consider the application of a filter to a test dataset with $n_l$ legitimate and $n_s$ Spam messages, resulting in $n_{l,s}$ and $n_{s,l}$ being incorrectly classified, respectively. In this case, it clearly follows that the number of correctly classified legitimate and Spam messages are given by $n_{l,l} = n_l - n_{l,s}$ and $n_{s,s} = n_s - n_{s,l}$ respectively.

In decision theory, two classes are labeled as positive (spam) and negative (legitimate), with the performance measures being the true positive $(TP = \frac{n_{s,s}}{n_s})$ and negative $(TN = \frac{n_{l,l}}{n_l})$ corresponding to the relative number of instances of each class that have been correctly classified. From these, the false positive and negative rates can be obtained $FP = 1 - TN$ and $FN = 1 - TP$.

### A. Confusion Matrix

The confusion matrix is a table contains information about the number of false positives, false negatives, true positives, and true negatives. Performance of algorithm is evaluated using the data in the matrix [14], [15].

Table I shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study:

- TP is the number of correct predictions that an instance is positive,
- FP is the number of incorrect predictions that an instance is positive,
- TN is the number of correct of predictions that an instance negative,
- FN is the number of incorrect predictions that an instances negative [16].

TABLE I
Table Type Styles

| a | b | Classe |
|---|---|---|
| TP (True Positives) | FP (False Positives) | a=0 |
| FN (False Negatives) | TN (True Negatives) | b=1 |

### B. Several Standard Terms Defined

$$\text{Precision} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FP} = TPR \tag{4}$$

$$\text{F.Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

$$k = \frac{P_0 - P_e}{1 - P_e} \tag{6}$$

The last parameter "kappa" is the amount of agreement correct by the agreement expected by chance with $P_0$: the proportion of the sample on which both judges agree and

$$P_e = \frac{\sum_i p_i p_{.i}}{n^2} \tag{7}$$

where: $p_i$: Sum of elements of the line $i$; $p_{.i}$ sum of elements of the column $i$; $n$ size of sample.

## V. Experimental Work and Results

In this section, we compare the classification accuracy results of alternating decision tree algorithms ADTree and classification algorithm Naive Bayes in the case of spam classification.

The dataset used for these tests named Spambase data set was created by Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt at Hewlett-Packard Labs [17]. It includes 4601 observations corresponding to email messages, 1813 of which are spam (39.4%) and Non -Spam are 2788 (60.6%). From the original email messages, 58 different attributes [18] UCI Machine Learning collection of spam e-mails came from their postmaster and individuals who had filed spam. Their collection of non-spam emails came from personal e-mails.

We have converted the Spambase.data data set into the Spambase.arff (Attribute Relation File Format).

The structure of Spambase.arff takes the following form:

```
@relation spambase
@attribute word_freq_make REAL
…..
@attribute class {0, 1}
@data
```

Next, we have eliminated the unnecessary attributes. We have eliminated two attributes *capital_run_length_longest Numeric* and *capital_run_length_total Numeric.*

The Spambase.arff has an attribute *class Nominal* denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail.

*A. Results for Classification Using Naive Bayes Algorithm*

Algorithm Naive Bayes is applied on the Spambase.arff dataset and the confusion matrix is generated for class gender having two possible values 0 or 1.

TABLE II
CONFUSION MATRIX OF NAÏVE BAYES

| a | b | classified as |
|---|---|---|
| 1920 | 868 | a=0 |
| 78 | 1735 | b=1 |

For above confusion matrix, true positives for class a = 0 is 1920 while false positives is 868 whereas, for class b = 1, true positives is 78 and false positives is 1735. diagonal elements of matrix 1920+1735 = 3655 represents the correct instances classified and other elements 78+868 = 946 represents the incorrect instances.

- TP rate for class a = 1920/(1920+868) = 0.689
- FP rate for class a = 78/(78+1735) = 0.043
- TP rate for class b = 1735/(1735+78) = 0.957
- FP rate for class b = 868/(868+1920) = 0.311
- Precision for class a = 1920/(1920+78) = 0.961
- Precision for class b = 1735/(1735+868) = 0.667
- F-measure for class a = 2*0.961*0.689/(0.961+0.689)= 0.802
- F-measure for class b = 2*0.667*0.957/(0.667+0.957)= 0.786

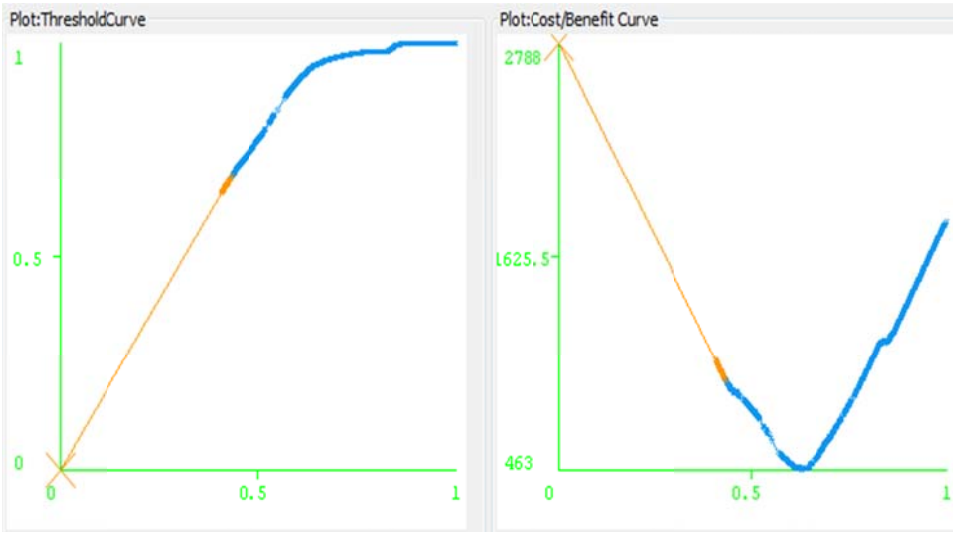


Fig. 2 Cost/Benefit Analysis of Naive Bayes (class = 0)

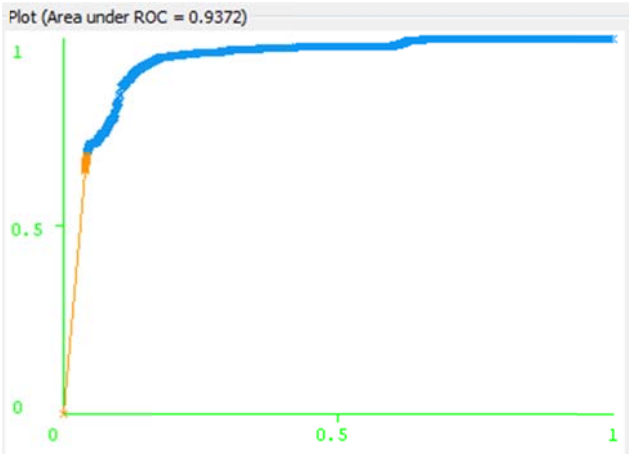


Fig. 3 Classifier Visualze: ThresholdCurve – bayes.NaiveBayes (class=0)

*B. Results for Classification Using ADTree Algorithm*

Algorithm ADTree is applied on the Spambase.arff dataset and the confusion matrix is generated for class gender having two possible values 0 or 1.

TABLE III
CONFUSION MATRIX OF ADTREE

| a | b | classified as |
|---|---|---|
| 2616 | 172 | a=0 |
| 162 | 1651 | b=1 |

For above confusion matrix, true positives for class a = 0 is 2616 while false positives is 172 whereas, for class b = 1, true positives is 162 and false positives is 1651. Diagonal elements of matrix 2616+1651 = 4267 represents the correct instances classified and other elements 162+72 = 334 represents the incorrect instances.

- TP rate for class a = 2616/(2616+172) = 0.938
- FP rate for class a = 162/(162+1651) = 0.089
- TP rate for class b = 1651/(162+1651) = 0.911
- FP rate for class b = 172/(172+2616) = 0.062
- Precision for class a = 2616/(2616+162) = 0.942
- Precision for class b = 1651/(1651+172) = 0.906
- F-measure for class a = 2*0.942*0.938/(0.942+0.938)= 0.940

- F-measure for class b = 2*0.906*0.911/(0.906+0.911)= 0.908

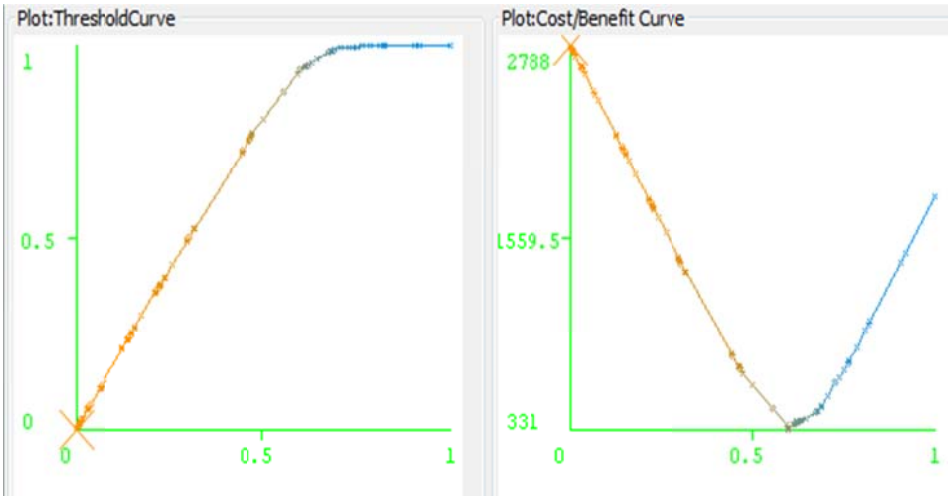


Fig. 4 Cost/Benefit Analysis of ADTree (class = 0)

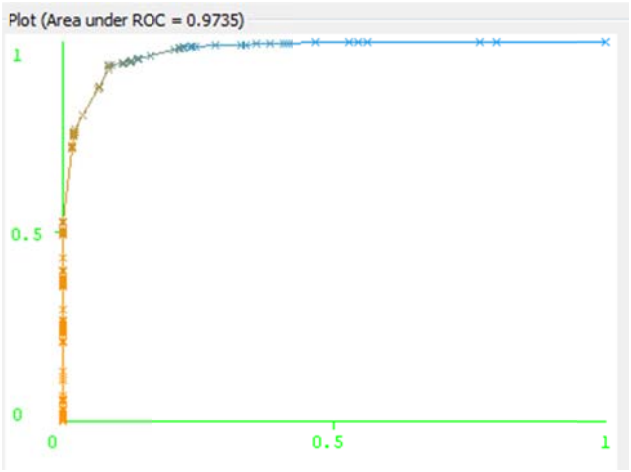


Fig. 5 Classifier Visualze: ThresholdCurve – trees.ADTRee (class=0)

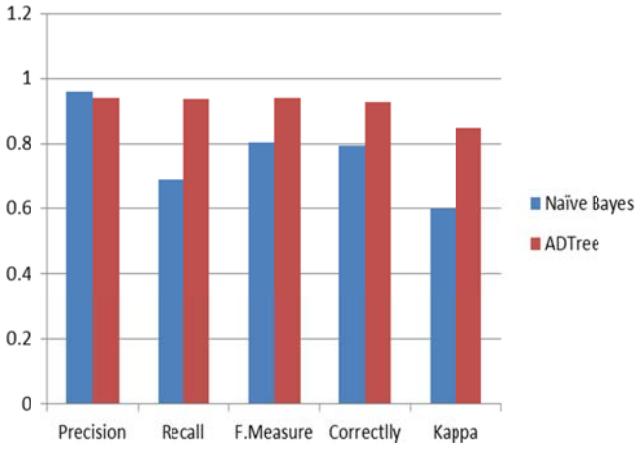


Fig. 6 Accuracy of the Spam classification algorithm

### *C. Comparison of Results*

From above experimental work we can conclude that correct instances generated by Naive Bayes are 3655 and ADTree are 4267, as well as performance evolution on the basis of Spambase dataset are shown in Table IV.

TABLE IV
DETAILED PRECISION CLASS ON TRAINING SET

| Algorithms | Precision | Recall | F.Measure | Correctly | Kappa |
|---|---|---|---|---|---|
| Naïve Bayes | 0.961 | 0.689 | 0.802 | 0.794 | 0.599 |
| ADTree | 0.942 | 0.938 | 0.940 | 0.927 | 0.848 |

## VI. CONCLUSION AND FUTURE WORK

In this paper, we compare two known methods: ADTree and Naïve Bayes algorithms using Weka tool for spam detection. The experiments results shown in the study are oriented classification accuracy and cost analysis.

ADTree gives more classification accuracy for class in Spambase data set having two values Yes and No.

Through the test results we obtained the performance evaluation value (Recall, F-Measure, Correctly and Kappa statistic) of the ADTree higher than Naïve Bayes. This proves that ADTree is better than the Naive bayes.

By reducing the number of attributes, we showed that ADTree Method is deemed appropriate for filtering SPAM in real time.

In the future work we want to combine statistical methods with decision tree for diagnosis and detection of cancer.

## REFERENCES

[1] S. M. Weiss and N. Indurkhya. Predictive data mining practical guide. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.
[2] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition and productivity. Technical report, McKinsey Global Institute, May 2011.
[3] Jiban K Pal , Usefulness and applications of data mining in extra cting information from different perspectives, Annals of Library and Information Studies, Vol. 58, March 2011, pp. 7-16
[4] http://www.radicati.com/wp/wp-content/uploads/2012/10/Email Market -2012-2016-Executive-Summary.pdf.
[5] Data Mining/ Data Warehousing Mosud Y. Olumoye Lagos State Polytechnic, S.P.T.S.A. & Director of Operations, Fiatcom Nig. Ltd. Nigeria.
[6] G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
[7] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.
[8] Sita Gupta, Vinod Todwal, Web Data Mining & Applications, nternational Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 –8958, Volume-1, Issue-3, February 2012.
[9] Data mining classification Fabriciovoznika Leonardoviana
[10] George Dimitoglou, James A. Adams, and Carol M. Jim, Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability.
[11] Seongwook Youn, Dennis McLeod, A Comparative Study for Email Classification.
[12] Yoav Freund and Llew Mason. The Alternating Decision Tree Algorithm. Proceedings of the 16th International Conference on Machine Learning, pages 124-133 (1999).
[13] Bernhard Pfahringer, Geoffrey Holmes and Richard Kirkby, Optimizing the Induction of Alternating Decision Trees, Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. 2001, pp. 477-487.
[14] Anshul Goyal and Rajni Mehta, Performance Comparison of Naïve Bayes and J48 Classification Algorithms, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012).
[15] Tina R. Patil, Mrs. S.S. Sherekar, Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification, Internationl Jpournal of Computer Science And Applications, Vol. 6, No.2, Apr 2013.
[16] Xiang yang Li, Nong Ye, A Supervised Clustering and Classification Algorithm for Mining Data With Mixed Variables, IEEE Transactions on Systems, man, and Cybernetics, Vol. 36, No. 2, 2006, pp. 396-406.
[17] https://archive.ics.uci.edu/ml/datasets/Spambase (Accessed online on January 2016).
[18] http://archive.ics.uci.edu/ml/. (Accessed online on January 2016).