ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Comparison Study on Performance Analysis of Data Mining Algorithms in Classification of Local Area News Dataset using WEKA Tool

G.Kesavaraj*1, Dr.S.Sukumaran2

[*1]PhD Research Scholar, Manonmaniam Sundaranar University, Assistant Professor, Vivekanandha College of Arts and Sciences For women, Thiruchencode Tamilnadu, India

[2] Associate Professor, Department of Computer science, Erode Arts Science College (Autonomous), Erode-9, Tamilnadu, India

kesavaraj2020@gmail.com

Abstract

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), [1] a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It is commonly used in marketing, surveillance, fraud detection, scientific discovery and now gaining wide way in social networking. Anything and everything on the Internet is fair game for extreme data mining practices. Social media covers all aspects of the social side of the internet that allow us to get contact and carve up information with others as well as intermingle with any number of people in any place in the world. This paper uses the dataset “Local News Survey” from Pew Research Center. The focus of the research is towards exploration on impact of the internet on Local News activities using Data Mining Techniques. The original dataset contains 102 attributes which is very large and hence the essential attributes required for the analysis are selected by feature reduction method. The selected attributes were applied to Data Mining Classification Algorithms such as RndTree, ID3, K-NN, C4.5 and CS-MC4. The Error rates of various classification Algorithms were compared to bring out the best and effective Algorithm suitable for this dataset.

General Terms: Classification algorithm, Rnd Tree algorithms, Error rates

Keywords: KDD, data mining, online surveys.

Introduction

Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. Data Mining involves an incorporation of techniques from multiple disciplines such as database and data warehouse technologies, statistics, machine learning, pattern recognition, neural networks and data visualization. A number of Algorithms have been developed and implemented to dig out information and discern knowledge patterns that may be constructive for decision support. Once these patterns are extracted they can be used for automatic classification of case mixes. Classification and prediction are the techniques used to make out important data classes and predict probable trends. Anything and everything on the Internet is fare game for extreme data mining practices. Social media covers all aspects of the social side of the internet that allow us to get contact and carve up information with others as well as interact with any number of people in any place in the world.

D. E. Brown, V. Corruble, and C. L. Pittard compared decision tree classifiers with back propagation neural networks for multimodal classification problems. J. Catlett has explained how knowledge patterns can be generated from large databases. M. James in his work describes the various classification algorithms. T. Cover and P. Hart performed classification using K-NN and proved its accuracy.

The dataset used in this paper is from “Local News Survey 2011 “obtained from a new national survey by the Princeton Survey Research Associates International. It is s nationally representative phone survey of 1,005 adults (ages 18+) was taken August 2-5, 2012. It was conducted in English on landline and cell

ISSN: 2277-9655
Impact Factor: 1.852

phones. The sample contained 799 internet users, who were asked questions about their online activities. The margin of error for the full sample is ± 3.7 percentage points. The margin of error for the internet sample is ±3.8 percentage points.

About Survey:

This Survey is based on the findings of a survey on Americans' use of the Internet. The results in this report are based on data from telephone interviews conducted by Princeton Survey Research Associates International from January 12 to 25, 2011, among a sample of 2,251 adults, age 18 and older. Telephone interviews were conducted in English and Spanish by landline (1,501) and cell phone (750, including 332 without a landline phone). For results based on the total sample, one can say with 95% confidence that the error attributable to sampling is plus or minus 2.4 percentage points. For results based Internet users (n=1,762), the margin of sampling error is plus or minus 2.7 percentage points. In addition to sampling error, question wording and practical difficulties in conducting telephone surveys may introduce some error or bias into the findings of opinion polls.

A combination of landline and cellular random digit dial (RDD) samples was used to represent all adults in the continental United States who have access to either a landline or cellular telephone. Both samples were provided by Survey Sampling International, LLC (SSI) according to PSRAI specifications. Numbers for the landline sample were selected with probabilities in proportion to their share of listed telephone households from active blocks (area code + exchange + two-digit block number) that contained three or more residential directory listings. The cellular sample was not list-assisted, but was drawn through a systematic sampling from dedicated wireless 100-blocks and shared service 100-blocks with no directory-listed landline numbers.

New sample was released daily and was kept in the field for at least five days. The sample was released in replicates, which are representative subsamples of the larger population. This ensures that complete call procedures were followed for the entire sample. At least 7 attempts were made to complete an interview at a sampled telephone number. The calls were staggered over times of day and days of the week to maximize the chances of making contact with a potential respondent. Each number received at least one daytime call in an attempt to find someone available. For the landline sample, interviewers asked to speak with the youngest adult male or female currently at home based on a random rotation. If no male/female was available, interviewers asked to speak with the youngest adult of the other gender. For the cellular sample, interviews were conducted with the person who answered the phone. Interviewers verified that the person was an adult and in a safe place before administering the survey. Cellular sample respondents were offered a post-paid cash incentive for their participation. All interviews completed on any given day were considered to be the final sample for that day.

Weighting is generally used in survey analysis to compensate for sample designs and patterns of non-response that might bias results. A two-stage weighting procedure was used to weight this dual-frame sample. The first-stage weight is the product of two adjustments made to the data – a Probability of Selection Adjustment (PSA) and a Phone Use Adjustment (PUA). The PSA corrects for the fact that respondents in the landline sample have different probabilities of being sampled depending on how many adults live in the household. The PUA corrects for the overlapping landline and cellular sample frames.

The second stage of weighting balances sample demographics to population parameters. The sample is balanced by form to match national population parameters for sex, age, education, race, Hispanic origin, region (U.S. Census definitions), population density, and telephone usage. The White, non-Hispanic subgroup is also balanced on age, education and region.

Organization of the Paper

The paper is organized as follows: Section 2 gives the idea about the dataset and WEKA tool used in this research categorization which is used for this research and Section 3 defines the proposed system and its functionality. Analysis and output of the system are presented in Section 4 and finally, Section 5 gives the conclusion of the research paper.

Data Set & Weka

Data Set Description

The dataset used in this paper is from “Local News Survey 2011” obtained from a new national survey by the Pew Research Center. This report is based on the findings of a survey on Americans' use of the Internet. The Dataset includes 162 attributes with 2251 records. The attributes were based on the questions posed towards the people. Some of the Sample questions in the survey included in the following Table 1 :

Q.No	**Sample Questions**
1	Are you under 18 years old, OR are you 18 or older?
2	Overall, how would you rate YOUR COMMUNITY as a place to live?
3	How much impact do you think people like you can have in making your community a better place to live — a big impact, a moderate impact, a small impact, or no impact at all?

ISSN: 2277-9655
Impact Factor: 1.852

4	In general…How much do you enjoy keeping up with the news – a lot, some, not much, or not at all?
5	Which of the following two statements best describes you…?
6	Thinking about ALL of the local news and information sources you use…How well do these sources give you the information you need? Would you say they cover…
7	If the only way to get full access to your local newspaper ONLINE on your computer, cell phone or other device was to pay a [FORM A READ: $10/FORM B READ: $5] monthly subscription fee, would you pay it or not?
8	Next I am going to read you some different sources where you might or might not get information about your local community. Please tell me how often, if ever, you use each source. (First/Next), how about…(INSERT IN ORDER) – [READ FOR FIRST ITEM THEN AS NECESSARY: do you get local information from this source every day, several times a week, several times a month, less often, or never?]

Table 1: Sample Questions

The original dataset is very vast with 102 attributes. To begin with, it is categorized into subsets for analysis of s Algorithms in Data Mining which is shown in the Table 1.

WEKA

WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves). It also has a general API embedded WEKA, like any other library, in any applications to such things as automated server-side data-mining tasks. The following Figure 2 shows WEKA software Explorer screen.

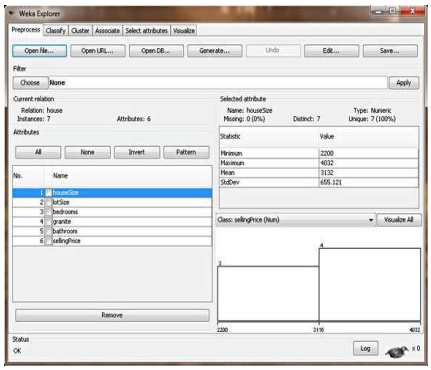


Fig. 2. WEKA software Explorer screen display.

Proposed System Model

This section deals with the architecture of the proposed system model which is shown in Figure 1. The subsets of the original dataset as described in Table 1 are considered for further analysis of Classification Algorithms.

It includes the following phases:

1. Data Cleaning (Handling missing Values).
2. Data Pre-processing (i.e., Applying Transformation and Feature Reduction).
3. Applying Classification Algorithms using WEKA tool.
4. Analysis of error rates produced by Algorithms.
5. Identifying the best Algorithm for this dataset.

Data Cleaning

Data Cleaning is also referred to as *data scrubbing*, the act of detecting and removing and/or correcting a databases dirty data (i.e., data that is incorrect, out-of-date, redundant, incomplete, or formatted incorrectly). The goal of data cleansing is not just to clean up the data in a database but also to bring consistency to different sets of data that have been merged from separate databases. The process stops when there is no variable to remove.

Data Transformation

Data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system. Data transformation can be divided into two Steps: 1. Data mapping maps data elements from the source data system to the destination data system and captures any transformation that must occurred. Steps: 2. The Code generation that creates the actual transformation program. This process is stops when Data is available to usable form.

Feature Reduction

After Transformation process, some preprocessing of the data is to be carried out to proceed further. Feature Reduction is one of the preprocessing techniques. In this phase the important features required to implement the Classification Algorithm are identified. By Feature Reduction, the model complexity is reduced and it is easier to interpret. Moreover, the attenuation of the variables to collect is an advantage during the deployment of the model. In some cases, the variable selection enables to improve the model accuracy. Manual selection by an expert domain is certainly the best approach. But because the number of candidate descriptors is often large, it is not always possible in practice. so, it must select automatically the best variables. It is also use the automatic process as a preliminary approach in order to filter out the really irrelevant attributes. The various feature selection Algorithms that we were tried includes:

ISSN: 2277-9655
Impact Factor: 1.852

Feature Ranking:
This Algorithm ranks the attributes based on their relevance. A cutting rule enables to select a subset of these attributes. It is a supervised Algorithm; we must define the discrete target attribute. This approach does not take into consideration the redundancy of the input attributes.

Relieff Filtering:

This is a supervised Algorithm which will not consider the redundancy of the input attributes. At least two attributes must be available and the target attribute must be discrete. [3]

Fast Correlation based Filtering (FCBF):

It is a supervised feature selection Algorithm based upon a filtering approach i.e., processes the selection independently from the learning Algorithm. This Algorithm, unlike the ranking approaches, takes into consideration the redundancy of the input attributes.

Fisher Filtering:

It is a supervised feature selection Algorithms based upon a filtering approach i.e., processes the selection independently from the learning Algorithm. This component ranks the inputs attributes according to their relevance. It is a supervised Algorithm; we must define the discrete target attribute. This approach does not take into consideration the redundancy of the input attributes.

Stepwise discriminant:

Step disc is always associated to discriminant .We implement the FORWARD and the BACKWARD strategies in WEKA. In the FORWARD approach, at each step, we determine the variable that really contributes to the discrimination between the groups. We add this variable if its contribution is significant. The process stops when there is no attribute to add in the model. In the BACKWARD approach, we begin with the complete model with all descriptors. We search which is the less relevant variable. We remove this variable if the removing does not significantly damage the discrimination between groups. The process stops when there is no variable to remove.

Correlation based Feature Selection (CFS):

It is a supervised feature selection Algorithm based upon a filtering approach .i.e. processes the selection independently from the learning Algorithm. This Algorithm unlike the ranking approaches, takes into consideration the redundancy of the input attributes. [3]

MIFS Feature Filtering:

It is a supervised feature selection Algorithm based upon a filtering approach. .i.e. processes the selection independently from the learning Algorithm. This Algorithm unlike the ranking approaches, takes into consideration the redundancy of the input attributes.

Multivalued Oblivious Decision Tree Feature Selection (MOD Tree):

It is a supervised feature selection Algorithm based upon a filtering approach. .i.e. processes the selection independently from the learning Algorithm. This Algorithm unlike the ranking approaches, takes into consideration the redundancy of the input attributes.

Runs Filtering:

It is a supervised feature selection Algorithm based upon a filtering approach. .i.e. processes the selection independently from the learning Algorithm. This component ranks the input attributes according to their relevance. [3]

Classification Algorithms

The goal of Classification is to build a set of models that can correctly foresee the class of the different objects. Classification is a two-step process, 1. Build model using training data. Every object of the data must be pre-classified i.e. its class label must be known. 2. The model generated in the preceding step is tested by assigning class labels to data objects in a test dataset. The test data may be different from the training data. Every element of the test data is also reclassified in advance. The accuracy of the model is determined by comparing true class labels in the testing set with those assigned by the model. The input to these methods is a set of objects (i.e., training data), the classes which these objects belong to (i.e., dependent variables), and a set of variables describing different characteristics of the objects (i.e., independent variables). The key advantage of supervised learning methods over unsupervised methods (for example, clustering) is that by having an explicit knowledge of the classes the different objects belong to these Algorithms can perform an effective feature selection if that leads to better prediction accuracy. The following are brief outline of some Classification Algorithms that had been used in data mining and machine learning area and used as base Algorithms in this research.

KNN Algorithms

KNN is an *non parametric lazy learning* algorithm. That is a pretty concise statement. When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world , most of the practical data does not obey the typical theoretical assumptions made (eg Gaussian mixtures, linearly separable etc) . Non parametric algorithms like KNN come to the rescue here.

It is a lazy algorithm. It does not use the training data points to do any *generalization*. In other words, there is *no explicit training phase* or it is very minimal. This means the training phase is pretty fast. It keeps all the training data. More exactly, all the training data is

ISSN: 2277-9655
Impact Factor: 1.852

needed during the testing phase. (Well this is an exaggeration, but not far from truth). This is in contrast to other techniques like SVM where you can discard all non support vectors without any problem. Most of the lazy algorithms – especially KNN – make decision based on the entire training data set (in the best case a subset of them). The dichotomy is pretty obvious here – There is a non existent or minimal training phase but a costly testing phase. The cost is in terms of both time and memory. More time might be needed as in the worst case; all data points might take point in decision. More memory is needed as we need to store all training data.

KNN Algorithms

Steps 1: Find the k closest training points (small kxi - x0k according to some metric, for ex. euclidean, manhattan, etc.)
Steps 2: Predicted class: majority vote
Steps 3: Predicted value: average weighted by inverse distance

Figure: 4 KNN algorithms

Create a root node for the tree
If all examples are positive, Return the single-node tree Root, with label = +.
If all examples are negative, Return the single-node tree Root, with label = -.
If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
Otherwise Begin
A = The Attribute that best classifies examples.
Decision Tree attribute for Root = A.
For each possible value, , of A,
Add a new tree branch below Root, corresponding to the test A = . Let Examples() be the subset of examples that have the value for A
If Examples() is empty
Then below this new branch add a leaf node with label = most common target value in the examples
Else below this new branch add the subtree ID3 (Examples(), Target_Attribute, Attributes – {A})
End
Return Root

ID3 (Iterative Dichotomiser 3) Algorithm

ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree. ID3 is the precursor to the C4.5 algorithm.
The ID3 algorithm can be summarized as follows:
Take all unused attributes and count their entropy concerning test samples
Choose attribute for which entropy is minimum (or, equivalently, information gain is maximum)
Make node containing that attribute
The algorithm is as follows:

C4.5 algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.
In pseudocode, the general algorithm for building decision trees is
Check for base cases
For each attribute a
Find the normalized information gain from splitting on a
Let a_best be the attribute with the highest normalized information gain
Create a decision node that splits on a_best
Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node

RndTree (Random Forest):

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).
Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

Decision List

The decision list induction is an ordered list of conjunctive rules [12]. It can handle a multi class problem. The obtained classifier gives an ordered set of rules.

6 Naïve Bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification

ISSN: 2277-9655
Impact Factor: 1.852

Analysis and Results

This section shows the analysis after executing various Classification Algorithms as per the requirements and explores the results of the same. The whole experiment is carried out with the Data Mining tool WEKA. The analysis of Feature Reduction technique is described in section 4.1 and the analysis of execution of the Classification Algorithm is described in section 4.2.

Analysis of Feature Reduction

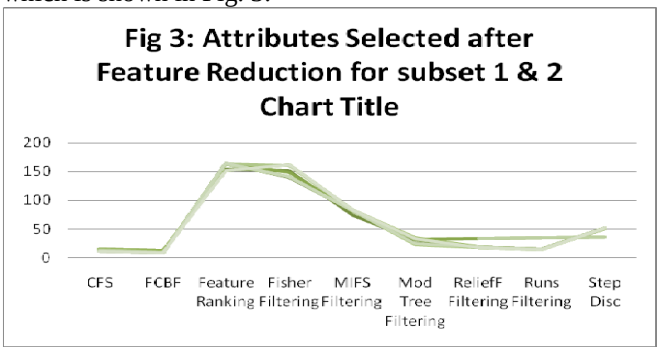
The features selected by feature reduction technique are chosen as input attributes with necessary class variable as the target attribute and various classification Algorithms were executed for all selected features one by one. The total number of attributes in the original dataset is 162. After performing feature reduction for the required subsets as shown in Table 1, important attributes were selected whose counts are shown in Table 3 & Table 4.

Table 3. Attributes selected after Feature Reduction for subset 1 & 2

Feature Selection Algorithms	Internet users	Mobile Phone Users
CFS	10	15
FCBF	8	8
Feature Ranking	165	165
Fisher Filtering	140	140
MIFS Filtering	82	82
Mod Tree Filtering	24	24
ReliefF Filtering	16	18
Runs Filtering	14	16
Step Disc	52	55

It does not imply that higher the number of attributes selected higher the accuracy of the classification algorithm. Even if less number of attributes were used, the attributes selected should be highly relevant for the target attribute or class attribute. For subset 1, the features selected by Feature ranking gave good results. For subset 2, reliefF Filtering produced good results. For subset 3, with a set of nine questions, same feature reduction Algorithms were applied and relevant attributes were identified and the counts of attributes selected are shown in Table 4.

Different algorithms gave different attributes and the best is selected for every survey question separately and necessary graph is drawn for the same, a sample of which is shown in Fig. 3.



	1	2	3	4	5	6	7	8	9	10
CFS	10	14	14	14	14	14	14	13	10	11
FCBF	8	12	9	10	12	9	10	9	8	9
Feature Ranking	165	154	163	164	151	163	164	163	165	151
Fisher Filtering	140	152	141	150	162	141	162	141	141	162
MIFS Filtering	82	74	82	81	83	82	83	82	82	83
Mod Tree Filtering	24	30	33	32	34	24	33	32	24	34
ReliefF Filtering	18	19	18	34	18	18	18	34	18	18
Runs Filtering	14	14	15	35	14	14	15	35	14	14
Step Disc	52	52	51	36	52	52	51	36	52	52

Analysis of Classification Algorithm

In this section we present a comparative study of various data mining classification algorithms on the Dataset “Local Area News”. For subset 1, the features selected by Feature ranking gave good results. For subset 2, reliefF Filtering produced good results.

The features selected by feature reduction technique are chosen as input attributes with necessary class variable as the target attribute and various classification Algorithms were executed for all selected features one by one. For subset 1 & 2, relevant attributes identified by feature reduction are executed by various Classification Algorithm and different error rates were identified and mentioned in the Table 5.

Error Rates of Classification Algorithms for sub set 1&2

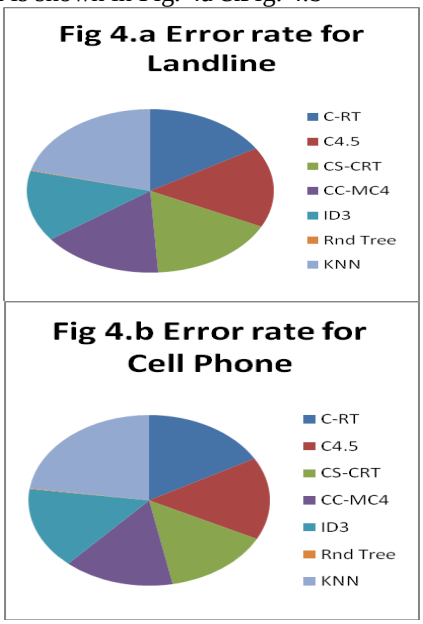
Table 5. Error rates of Classification Algorithms (After Execution)

Algorithm	Error rate for Landline	Error rate for Cell Phone
C-RT	0.1458	0.1402
C4.5	0.1422	0.1300
CS-CRT	0.1500	0.1200
CC-MC4	0.1423	0.1233
ID3	0.1256	0.1285
Rnd Tree	0.0012	0.0012
KNN	0.1872	0.1892

ISSN: 2277-9655
Impact Factor: 1.852

Navie bayces	0.1956	0.1658

From Table 5, it is clear that the error rate generated by Rnd Tree Algorithm is very less compared to all other Algorithms. The misclassifications identified were very less. A Graph drawn for the error rates after executing the Algorithm for the attributes selected by Feature reduction is shown in Fig. 4a &Fig. 4.b



A confusion Matrix is obtained. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). A sample Confusion Matrix for RnDTree Classification Algorithm is shown in

	a	b	c	d	Sum
a	1125	0	0	0	1125
b	0	456	0	0	456
c	0	0	125	0	125
d	0	0	0	74	74
sum	1125	456	125	74	1780

Fig. 5 A Sample Confusion Matrix for RnD Tree Algorithm for subset 2

Figure 5. In the Figure 5, n, d, s, N and r are various identifiers and the descriptions are shown in the Table 7.

Table 6. Description of Confusion Matrix.

a	Cell phone user
b	Not a cell phone user
c	Internet user
d	Refused to answer

Similarly for subset 3 also different Algorithms were tried and the corresponding error rates for different survey questions are shown in the Table 7.

Table 7. Error rates of Classification Algorithms (After Execution) subset 3

Algorithm	Error rate for Landline	Error rate for Cell Phone
C-RT	0.1460	0.1457
C4.5	0.1467	0.1345
CS-CRT	0.1545	0.1225
CC-MC4	0.1343	0.1400
ID3	0.1456	0.1600
Rnd Tree	0.0024	0.0024
KNN	0.1872	0.1892
Navie bayce	0.1956	0.1954

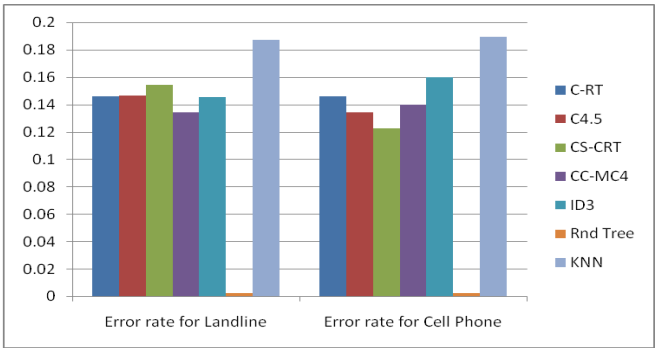


Figure 6. Error rates for various classification algorithms

From Figure 6 , we can infer that n, d, s and r have no misclassifications whereas N has one misclassification where it has been identified as n. After analysis of the results it is clear that the Classification Algorithm RndTree gave lesser error rates when compared to other Classification Algorithms for this dataset and declared as best Algorithm with efficient as for as the dataset " Social Side of the Internet" is concerned.

Conclusion

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. Social network analysis applications have experienced tremendous advances within the last few years due in part to increasing trends towards users interacting with each other on the internet. There have been a large number of data mining

ISSN: 2277-9655
Impact Factor: 1.852

Algorithms rooted in these fields to perform different data analysis tasks. In this paper, the comparison on the performance of Data Mining Classification Algorithms was executed on the dataset "Local Area News". To start with the entire dataset is categorized into 3 subsets. The entire attribute set includes 102 attributes which is very vast and hence feature reduction is performed to identify the highly relevant attribute for the target variable. The selected attributes were given as input to various Data Mining Classification Algorithm and the error rates were analyzed and compared. From the results it is clear that in all the subsets considered for the research RndTree Algorithm produced less error rates when compared to all other Algorithms while executing with WEKA tool.

References

[1] http://www.cs.waikato.ac.nz/~ml/weka/.
[2] http://ieeexplore.ieee.org/A Comparable Studyemploying WEKA Clustering/ClassificationAlgorithms for Web Page Classification/IoannisCharalampopoulos, Ioannis Anagnostopoulos/ 2011,Page(s): 235 – 239.
[3] http://ieeexplore.ieee.org/Rule-Based ClassificationApproach for Railway Wagon HealthMonitoring/GM Shafiullah, A B M Shawkat Ali, AdamThompson, Peter J Wolfs/2010, Page(s): 1 – 7.
[4] http://www.kdnuggets.com/gpspubs/aimag-kddoverview-1996-Fayyad.pdf.
[5] http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf
[6] http://en.wikipedia.org/wiki/Data_mining.[7] Books: DATA MINING by Ian H. Witten & EibeFrank, Second Edition