MAT JOURNALS
Journal of Data Mining and Management
Volume 1 Issue 2

An Extensive Benchmark Experimental Evaluation of Methods for Multi Label Learning In R

J Uma Mahesh, N Chandrakanth, M Ravinder Reddy
Department of CSE, Geethanjali College of Engineering & Technology, Hyderabad, India
E-mail: umamaheshsjcet@gmail.com, chandra.nagulapalli@gmail.com, ravinder0512@gmail.com

Abstract

A smart product is one that is able to immingle with masses. Sensible merchandise does not seem to be solely easy merchandise, however, with a touch of cleverness supplemental to permit the user some flexibility operative. Smart product adapts to the context of the user and does not force the user to adapt to that. Sensible merchandise have a group of properties that creates them distinctive area unit self informative, self organizing, extensible, self property, device capabilities, practicality, integrity, user services, property. The client's ranking or priority whereas shopping for varied sensible merchandise area unit dynamical day by day as a result of advancements in technology and customer principally target the advanced options of the sensible merchandise they are shopping for. This paper principally shows, however, affectively sensible merchandise area unit utilized by the shoppers and area unit hierarchic based mostly upon their performance by exploitation R language and WEKA. By using R we can have a deep analysis over the various smart products and the user can be able to know the most widely purchased smart products according to their ranking. We can have deep analysis of smart products using data mining classification and prediction techniques such as J48, Random Forest machine learning algorithms in WEKA and R Language. R allows wide number of smart products data and analyzes with in limited resources. The WEKA and R language is opted to see the classification and prediction performances. Four measures (sensitivity, specificity, accuracy, F–measure) of performance here considered are based on confusion matrix/Error Matrix of R and WEKA, table of counts revealing the performance of each algorithm's confusion regarding the true classifications and predictions. The observation of all the four performance measures used to analyze the smart products use.

Keywords: *Decision tree, random forest, confusion, matrix, R part, rattle package, WEKA, R*

INTRODUCTION

Data Mining

Data mining is a process of extracting knowledge from massive volume of data. It refers to the way of finding significant and useful information from a data base. The knowledge which is extracted can include patterns, association rules and different trends. Data mining is not confined to any particular organization instead it has many techniques used to explore knowledge hidden in any data [1]. Different techniques used for digging the data out are artificial intelligence, statistical and numerical techniques and pattern recognition techniques. Using data mining techniques we examine the large pre-existing data bases in order to generate the new information. To build any classification model for predicting the data we perform some analysis in order to predict or classify the new data.

Data Mining Classifier Algorithms

J48 Algorithm

J48 classifier is a simple C4.5 decision tree for classification which creates a binary tree. The

decision tree approach is most useful in classification problem. Using this technique, a tree is constructed to model the classification process. Once the tree is constructed, it is applied to each tuple in the database and results in classification for that tuple

Algorithm J48

INPUT

Data //Training data

OUTPUT

Tree //Decision tree

DTBUILD (*Data)

{

Tree=φ;

Tree= Create root node and label with splitting attribute;

Tree= Add arc to root node for each split predicate and label;

For each arc do

Data= Database created by applying splitting predicate to Data;

If stopping point reached this path, then

Tree= create leaf node and label with appropriate class;

Else

Tree'= DTBUILD (Data);

Tree= add Tree to arc;

}

While building a tree, J48 ignores the missing values, i.e., the worth for that item may be foretold supported illustrious attribute values for the opposite records. The most plan is to divide the information into vary supported the attribute values for

that item that square measure found within the coaching sample. J48 permit classification via call trees or rules generated from them [2].

Random Forests

Random forests area unit an ensemble learning technique for classification, regression and different tasks, that operate by constructing a large variety of call trees at coaching time and outputting the category that is the mode of the categories (classification) or mean prediction (regression) of the individual trees. Random forests area unit correct for call trees' habit of over fitting to their coaching set. The algorithm for random forest was developed by Leo Breiman and Adele Cutler and "Random Forests" is their trademark. The method combines Breiman's "bagging" idea and the random selection of features, introduced autonomously by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variance [3–6].

The selection of random subset of features is an example of the random subspace method, which, in Ho's formulation, is a way to apply classification proposed by Eugene Kleinberg.

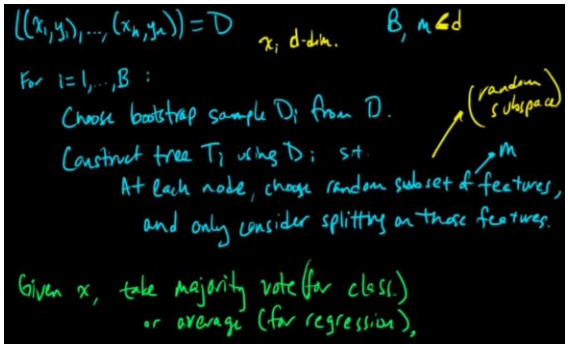


Fig. 1: *Random Forest Algorithm.*

RELATED WORKS

Case Study: Detecting Fraudulent Transactions (Refer Data Mining with R.)

This case study addresses associate degree mental representation of the overall drawback of detection uncommon observations of phenomena, that is, finding rare and quite completely different observations. The driving application should do with transactions of a group of merchandise that square measure reported by the salespeople of some company. The goal is to search out “strange” dealing reports which will indicate fraud tries by a number of the salespeople. The result of the information mining method can support posterior review activities by the corporate. Given the restricted quantity of resources which will be allotted to the present review activity, we wish to produce a form of fraud likelihood ranking as outcome of the method. These rankings ought to permit the corporate to use its

review resources in associate degree best approach. This general resource-bounded inspection activity is frequent in many fields, such as credit card transactions, tax declarations inspection, etc. This chapter addresses several new data mining tasks, namely, (1) outlier or anomaly detection, (2) clustering and (3) semi-supervised prediction models. By this case study we came to know how to apply several data mining tasks, clustering, semi-supervised prediction models that are helpful for clear overview. By this we proposed statistical computing by using R language and rattle package for predicting the usage of smart products [7, 8].

Brief Introduction about R and WEKA

R

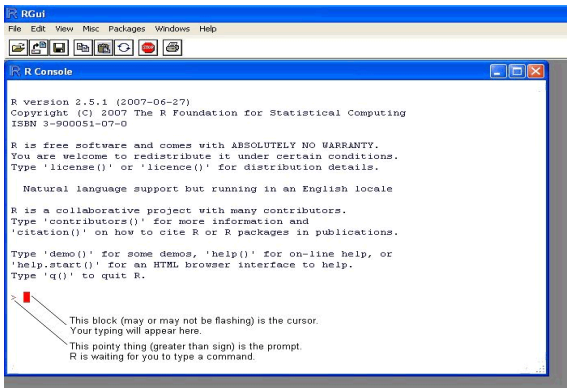


Fig 2: *Introduction about R.*

R Console Examples

R Session

After R is started, there is a console waiting for us to give some input. At the prompt (>), we can enter numbers and perform calculations.

> x <- 945

The effect of the previous instruction is thus to store the number 945 on an object named x. By entering the name of an object at the R prompt we can see its contents.

> x

[1] 945

The rather cryptic "[1]" in front of the number 945 can be read as "this line is showing values starting from the first element of the object."

We can also assign numerical expressions to an object. In this, the object will store the result of the expression

> z <- 5

> w <- z^2

> w

[1] 25

> i <- (z * 2 + 45)/2

> i

[1] 27.5

This means that "calculate whatever is given on the right side of the operator, and assign (store) the result of this calculation to the object whose name is given on the left side".

We can use R prompt as a kind of calculator:

```
> 1 + 2
[1] 3
```

Every object we create will stay in the computer memory until we delete it. We can know the list of objects currently in the memory by issuing the ls() or objects() command at the prompt.

```
> ls()
[1] "i" "w" "x" "y" "z"
```

If we do not need an object, we may free some memory space by removing it

```
> rm(y)
> rm(z, w, i)
```

Note: names in R are case sensitive, in which

Color and color are two different objects.

Functions

We can create a vector in R, using the c () function, it combines its arguments to form a vector. R functions are invoked by its name, which is followed by the parenthesis, with zero or more arguments.

```
> v <- c(4, 7, 23.5, 76.2, 80)
[1] 4.0 7.0 23.5 76.2 80.0
```

Comments

All text after the pound sign "#" within the same line is considered a comment.

```
> 1 + 1      # this is a comment
[1] 2
```

Packages

A set of add-on packages available for R at CRAN (Comprehensive R Archive Network). In the Windows version this is easily done through the "Packages" menu. After connecting computer to the Internet you should select the "Install package from packages option from R console menu. This option will present a list of the packages available at CRAN. You select the one you want, and R will download the package and self-install them on your system. If you want to know the packages currently installed in your computer, you can type the following command.

```
> installed.packages()
```

If you want to quit R you can issue the command q () at the prompt. If you want to save the current workspace we should answer yes only if we want to resume our current analysis at the point we are leaving it.

Getting Help

R provides extensive documentation. For example, entering? c or help(c) at the prompt gives documentation of the function c in R.

```
> help(c)
```

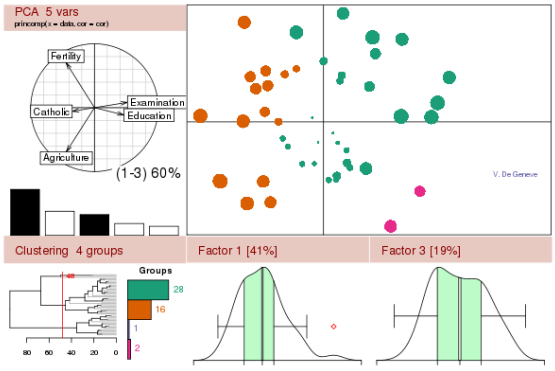


Fig. 3: *Statistics on Smart Products/Results.*

WEKA

WEKA is a collection of various machine algorithms for Data mining tasks. The algorithms can be directly applied to the data set or they can be taken through java code. WEKA contains tools for Pre-processing, Classification, Clustering, Association, Data visualization. WEKA is developed by University of Waikato sample picture.

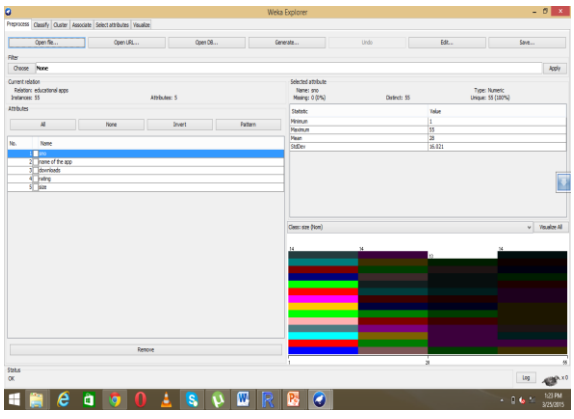


Fig. 4: *WEKA.*

SMART PRODUCTS DATASET DESCRIPTION AND EXPERIMENTAL SETUP

Smart Products

A smart product is one that is able to immingle with masses. Smart product is not solely straight forward product, however, with a touch of cleverness side to permit the user some flexibility operative.

Table 1: *Attributes of Smart Products Data Set.*

Name	**Description**
Model Name	Company Name
Front Camera	Front Facing Camera
Rear Camera	Backup Camera
Screen Size	Viewable Area
RAM	Storage
Internal Storage	Purpose of Storing the Data
Weight	Thing Weighs
Operating System	Program that Runs a Smart Phone
Bluetooth	Wireless Networking Technology
Wi-Fi	Connecting to the Internet
USB	Universal Serial Bus
Price	A Value that Purchase a Finite Quantity

Table 2: *Summary of Data Sets.*

Dataset	# Instances	# Attributes	# Classes
Smart Products	55	12	3

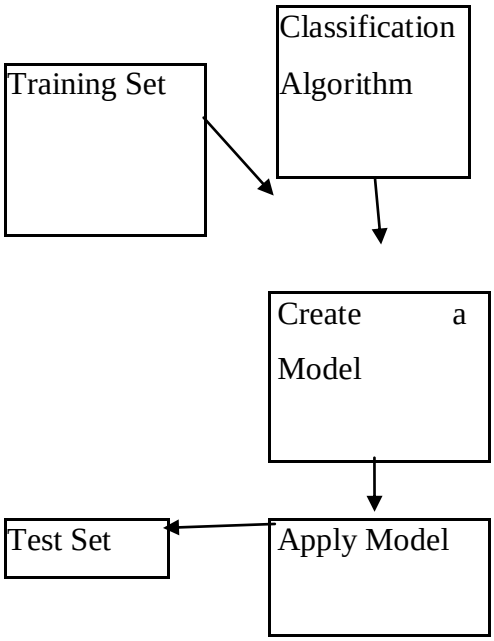


Fig. 5: *Overall Experimental Setup.*

>S.NO<-1:30

>modelname<-c("samsung galaxy grand prime","samsung galaxy grand","samsung galaxy grand duos","nokia lumia 520","nokia lumia 620","nokia lumia 720"," sony xperia m","sony xperia sp","sony xperia e4","sony xperia z3 dual","htc desire 626","htc nexus 9","htc one m8 eye","htc butterfly 2","huawei honor holly","huawei ascend gxi","huawei honor 3c 4g","huawei mediapad honor t1","lg icecream smart","lg l80 dual","micromax bolt d321","micromax canvas hueaq4000","blackberry classic","blackberry passport","iball slide 1044","dell venue 8 2014","hp 7 g2","motorola moto max","apple ipad mini3","yu yurekha")

>frontcamera<c(5,1.9,2,1.3,0.3,"no",0.3,2,2,5,5,1.6,5,5,2,2,5,0.3,2,0.3,"no",2,"no",2,2,2,0.3,2,2,5)

>rearcamera<c(8,8,8,6.7,5,5,5,8,5,20,13,8,13,13,8,8,8,5,8,5,5,8,"yes",13,2,5,2,21,5,13)

>batterycapacity<c(2600,2600,2100,2000,1300,1430,1750,2370,2300,3100,2000,6700,2600,2700,2000,3500,2300,4800,1700,2450,1800,3000,2515,3450,4000,4500,2000,3900,6471,2500)

>screensize<c(5,5.2,5,4.3,3.8,4,4,4.6,5,5.2,5,8.9,5,5,5,6,5,8,3.5,5,5,5,3.5,4.5,10,8,7,5.2,7.9,5.5)

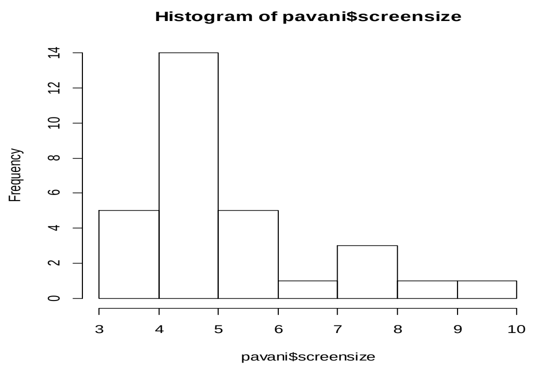
>RAM<c("1GB","1.5GB","1GB","512MB","512MB","512MB","1GB","1GB","1GB","3GB","1GB","2GB","2GB","2GB","1GB","1GB","1GB","1GB","1GB","1GB","512MB","1GB","2GB","3GB","1GB","1GB","1GB","3GB","1GB","2GB")

>INTERNALSTORAGE<c("8GB","8GB","8GB","8GB","8GB","8GB","4GB","4GB","8GB","16GB","16GB","32GB","16GB","16GB","16GB","8GB","8GB","8GB","8GB","4GB","4GB","8GB","16GB","2GB","8GB","16GB","8GB","64GB","16GB","16GB")

```
>WEIGHT<c(156,163,161,128,127,124,115,155,144,152,140,425,160,151,123,173,145,360,138,138,145,145,178,196,178,338,256,176,341,155)
> operatingsystem<-c("ANDROID 4.4","ANDROID 4.3","ANDROID 4.1","WINDOWS PHONE8","WINDOWS PHONE8","WINDOWS PHONE8","ANDROID 4.1","ANDROID 4.1","ANDROID 4.4","ANDROID 4.4.4","ANDROID 4.4","ANDROID 5.0","ANDROID 4.4","ANDROID 4.4","ANDROID 4.4.2","ANDROID 4.4","ANDROID 4.4","ANDROID 4.3","ANDROID 4.4","ANDROID 4.4","ANDROID 4.4.2","ANDROID 4.4.2","BLACKBERRY OS 10.3","BLACKBERRY OS 10.3","ANDROID 4.4","ANDROID 4.4","ANDROID 4.4.2","ANDROID 4.4.4","IOS8","ANDROID 4.4.4")
>bluetooth<c("yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes")
>wifi<c("yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes","yes")
>USB<-c("MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB","MICRO USB")
>price<c(13350,15203,17197,21533,13466,7999,8811,13999,9990,50990,11999,28900,38990,8999,6999,15999,9499,999,27999,15499,5499,10999,31990,49990,11999,18999,8799,53999,38900,8999)
```

Example:

```
>hist(pavani$screensize)
```




```
>plot(density(pavani$batterycapacity))
```

PROPOSED METHOD

The proposed method compares the traditional 2 machine learning classification and predictions algorithms in R language and WEKA. R provides many packages out of them Rattle, rpart, randon Forest packages are used for the analysis of the classification algorithms. WEKA provides many machine learning algorithms out of them J48 and Random Forest techniques are used for the analysis of the classification algorithms

Rattle Package

Rattle GUI may be a free and open supply code (GNU GPL V2) package providing a graphical program (GUI) for data processing mistreatment the R applied math artificial language. Rattle is employed in numerous things. Presently, fifteen numerous government departments in Australia and round the world use rattle in their data processing activities and as an applied math package. Rattle provides extended data processing practicality by exposing the facility of the R. Statistical computer code through a graphical interface. Rattle is additionally used as a coaching facility to find out the R computer code Language. A Log Code tab is gift that replicates the R code for any activity undertaken within the graphical user interface, which might be derived and glued. Rattle also can be used for applied math analysis, or model generation. Rattle permits the dataset to be divided into coaching, validation and testing. The dataset will be viewed and emended. associate in nursing possibility for grading associate in nursing external file is gift.

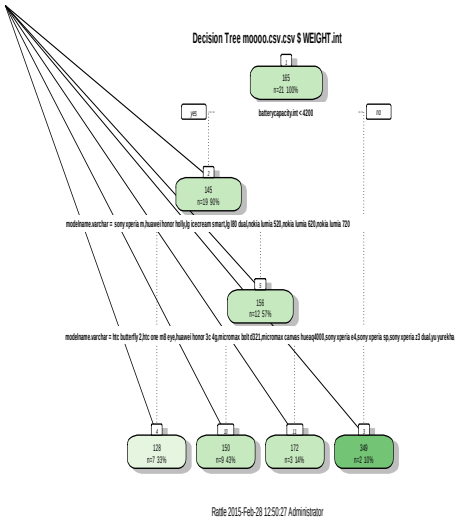
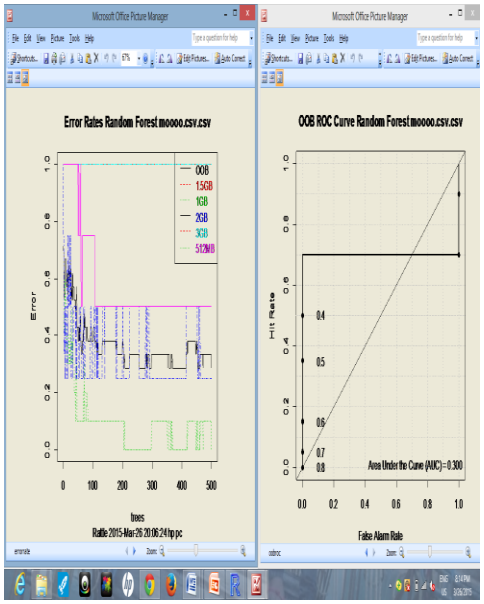


Fig. 6: *Tree Generated for Rpart.*



S.no	Classifier algorithm	WEKA	R
1	J48	70.12	71.23
2	RF	68.88	72.28

Fig. 7: *Graph Generated for Random Forest.*

RESULTS AND CONCLUSION

Measuring the Performance Accuracy

The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

Confusion matrix/Error Matrix

It is also known as a contingency table or an error matrix or table of confusion. Each column of the matrix represents the instances in a predicted class each row represents the instances in an actual class.

1. Accuracy=TP+TN/TP+TN+FP+FN
2. Sensitivity=TP/TP+FN
3. Specificity=TN/TN+FP
4. F-Measure=2 X precision X recall/precision + recall

   TP -> number of true positives TN -> number of true negatives N -> number of false negatives FP -> number of false positive

Sensitivity

Sensitivity is also referred as True positive rate, i.e., the proportion of positive tuples that are correctly identified.

Specificity

Specificity measures the proportion of negative tuples that are correctly identified.

F–Measure

The F–score is used as a single measure of performance of the test. It is the harmonic mean of precision and recall.

Performances by Metric

Smart Products Data

The following Table 3 shows the overall efficiency of the classifiers for smart products dataset in terms of 4 measures

Table 3: *Overall Efficiency of the Classifiers for Smart Products Dataset.*

Table	Performance of classifiers for smart products data							
Classifier	Sensitivity(%)		Specificity(%)		Accuracy(%)		F Measure	
	WEKA	R	WEKA	R	WEKA	R	WEKA	R
J48	54.6	52.4	77.2	73.3	70 .2	71.1	1.604	1.9334
PCA	51.9	61.2	76.1	78.9	68.8	69.10	1.625	1.2356

Figure 8 shows that the decision tree outperforms well for smart products data set.

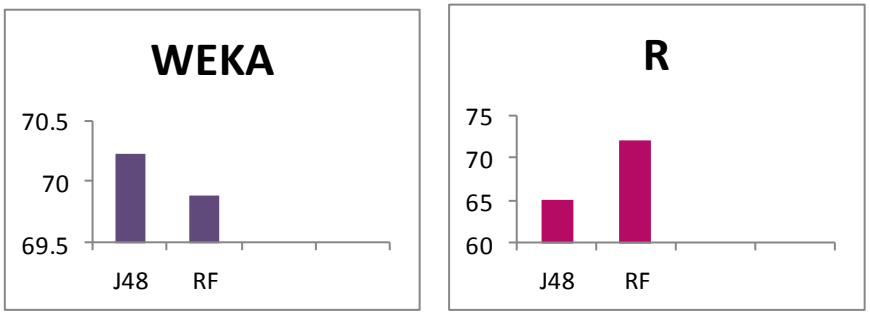


Fig. 8: *Decision Tree Outperforms Well for Smart Products Data Set.*

REFERENCES

1. Available at: http://en.wikipedia.org/wiki/Abalone.
2. Alexandros K, David. M, Kurt Hornik. Support vector machines in R. *Journal of Statistical Software*. 2006.
3. Blake, C. L., Merz, C. J. UCI repository of machine learning databases. http://www.ics.uci.edu/-_mlearn/MLRepository.html, 1998.
4. Breiman L., Friedman J.H., Olshen R.H., Stone C.J Classification and regression trees. Wadsworth and B Rooks, Monterey, CA; 1984.
5. Breiman, Leo. Random Forest–Random Features; 1999.
6. Domingo. P, M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*. 1997; 29 (2–3): 103–130p.
7. Graham J Williams. Rattle: A data mining GUI for R. *The R Journal*. 2009; 1/2.
8. Jiawei Han, Micheline Kam ber. Book on data mining: Concepts and Techniques. 2nd ed, Morga n Kaufmann Publishers; 2006.