Analysis of Machine Learning Algorithms using WEKA

Aaditya Desai Ph.D. Scholar, NMIMS University. TCET, Mumbai

ABSTRACT

The purpose of this paper is to conduct an experimental study of real world problems using the WEKA implementations of Machine Learning algorithms. It will mainly perform classification and comparison of relative performance of different algorithms under certain criteria.

General Terms

TreesJ48, TreesJ48graft, RandomTree, OneR, ZeroR, Decision Table, Naïve Bayes, Bayes Net, Naïve Bayes Simple, Bayes Updatable, Multilayer Perceptron, Logistic, RBF Network, Simple Logistic

Keywords

WEKA, Machine Learning

1. INTRODUCTION

WEKA is a collection of open source of many data mining and machine learning algorithms, including: pre-processing on data, classification, clustering, association rule extraction.[1] [2]

In this paper we have taken the real world problem of predicting whether it is going to rain or any other prediction of weather. Machine learning works on the concept of the way a human brain works the machine also uses logical steps to perform the decision or to predict an output.

2. Data Set

The Data Set consists of attributes related to weather conditions. These weather conditions are sunny, overcast and rainy. Temperature, humidity, windy will provide us the actual values to make a decision whether to play or not to play.

2.1 Description of attributes in the Data Set

Table 1.1 provides the description of the attributes in the data set. The selected attributes consists of discrete attribute type. Also Fig 1.1 shows the input format of the data set which is in ARFF form i.e. Attribute Relation File Format which is used as input to Weka.

outlook	temperatu	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

Table: 1.1 Weather.csv file

Dr. Sunil Rai Ph.D. Guide, NMIMS University.

Code for Weather1.arff:

@relation weather

@relation outlook{sunny,overcast,rainy}@attribute temperature real

@attribute humidity real

@attribute windy{TRUE,FALSE}

@attribute play{yes,no}

@data

sunny,85,85,FALSE,no

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

rainy,70,96,FALSE,yes

rainy,68,80,FALSE,yes

rainy,65,70,TRUE,no

overcast,64,65,TRUE,yes

sunny,72,95,FALSE,no

sunny,69,70,FALSE,yes

rainy,75,80,FALSE,yes

sunny,75,70,TRUE,yes

overcast,72,90,TRUE,yes

overcast,81,75,FALSE,yes

rainy,71,91,TRUE,no

3. Results and Discussion:

3.1 Implementation of Algorithms

Weka is chosen for implementation of algorithms. The objective of selecting this tool is to understand the basic concepts and also application of these algorithms in real time. Weka is helpful in learning the basic concepts of machine learning with different options and analyzes the output that is being produced.

Implementation Procedure Used in Weka:

The ARFF file is fed in to Weka and the classification algorithms are implemented as defined in the following steps:

a) In the Preprocess tab, Discretize filter is applied to discretize the attributes sunny, overcast and rainy.



Fig. 1.2 a)Choosing filter

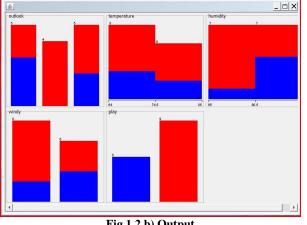


Fig.1.2 b) Output

b) In the Classify tab, choose the classification algorithm to be implemented and start the analysis to get results. In the WEKA experiment environment, the classifier at the top of the list in algorithms section is used as reference classifier. The table 1.2 below shows the classifiers that should be included. The *Attributes* field in the table states any change for the specified parameters of the classifier. Use 10-fold cross validation in the experiment.

Classifier	Attributes	Comments	Figure
TreesJ48	Autoutes	Without pruning, ID3	1.3
11668340	unprune=true		1.5
	_	without pruning	
TreesJ48gr	default	Uses pruning, C4.5	1.4
aft		algorithm: an	
		improvement over ID3.	
RandomTr	default	Random Tree	1.5
ee			
OneR	default	Rules OneR	1.6
ZeroR	default	Rules ZeroR	1.7
Decision	default	Rules Decision Table	1.8
Table			
Naïve	default	Naïve Bayes	1.9
Bayes			
Bayes Net	default	Bayes Net	1.10
Naïve	default	Naïve Bayes Simple	1.11
Bayes		i i j i i j i i i j i i i i i i i i i i	
Simple			
Bayes	default	Bayes Updatable	1.12
Updatable			
Multilayer	default	Function	1.13
Perceptron			
Logistic	default	Function Logistic	1.14
RBF	default	Function	1.15
Network			
Simple	default	Function Simple Logistic	1.16
Logistic		r c	

3.1.1 Classifier Algorithm

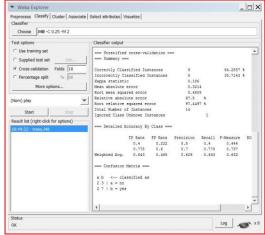


Fig. 1.3 J48

Results:

Time taken to build the model = 0 seconds Correctly classified instances (in %) = 64.299%Total number of instances = 14

 Weka Explorer 		- 0 ×
Preprocess Classify Cluster Associate Classifier	Select attributes Visualize	
Choose 348graft -C 0.25 -M 2		
Test options	Classifier output	
C Use training set C Supplied test set If Cross-validation Folds IO Percentage split № 66 More options	Scheme: werke.Classifiers.trees.J#Sgraft -C 0.25 -H 2 Relation: weather1 Istances: IS Attributes: S outlook temperature humidity	*
(Nom) play 🖉	windy play Test mode: 10-fold cross-validation Classifier model (full training set)	
Result (right-click for options) [1349:22 - traces.]48 [1353:13 - traces.]48 [1353:23 - traces.]48	Jähpraft prunet tree voltook # sunny 1 humidity << 75; yms (2.0) 1 humidity << 75; yms (2.0) voltook * rakny voltook * rakny 1 windy # TARE: pay (3.0) 1 windy # TARE: pay (3.0) 1 windy # TARE: pay (3.0)	
	Number of Leaves : 5 Size of the tree : 8 4	*

Fig. 1.4 J48graft

Results: Time taken to build the model = 0 seconds Correctly classified instances (in %) = 64.299%Total number of instances = 14

3.1.3 Classifier Algorithm RandomTree

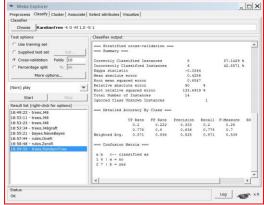


Fig.1.5 Random Tree

Results:

Time taken to build the model = 0 seconds Correctly classified instances (in %) = 57.143%Total number of instances = 14

3.1.4 Classifier Algorithm OneR

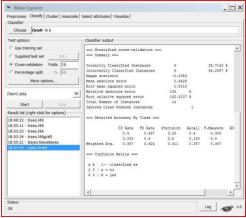


Fig. 1.6 One R

Results:

Time taken to build the model = 0 seconds Correctly classified instances (in %) = 35.714%Total number of instances = 14

3.1.5 Classifier Algorithm ZeroR

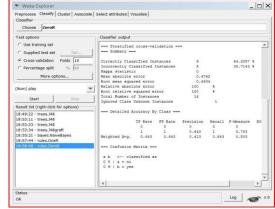


Fig. 1.7 ZeroR

Results: Time taken to build the model = 0 seconds Correctly classified instances (in %) = 64.286%Total number of instances = 14

3.1.6 Classifier Algorithm Decision Table

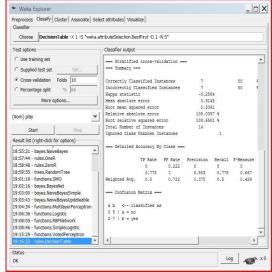


Fig. 1.8 Decision Table

Results:

Time taken to build the model = 0.01 seconds Correctly classified instances (in %) = 50%Total number of instances = 14

 Weka Explorer 										
Preprocess Classify Cluster Associate	Select attributes V	isualize								
Classifier										
Choose NaiveBayes										
Test options	Classifier output									
C Use training set	Stratified	cross-val	idation -	-						
C Supplied test set Set	Summary	-								
Cress-validation Folds 10	Correctly Clas	sified In:	stances	9		64.2857				
C Percentage split % 66	Incorrectly Cl	assified 1		5		35.7143	۹.			
	Kappa statisti			0.18						
More options	Mean absolute			0.44						
	Root mean squa			0.51 94.16						
(Nom) play 👻	Root relative			104.90						
	I Total Number of			14						
Start Stop	Ignored Class				1					
Result list (right-click for options)	1									
18:49:22 - trees.]48	Detailed 3	couracy Bj	Class ===							
18:53:11 - trees.J48		TP Rate	FP Rate	Precision	Bass11	F-Measure	80			
18:53:23 - trees.348		0.4	0.222	0.5	0.4	0.444	-			
18:53:34 - trees.J48graft		0.778		0.7	0.778					
18:55:21 - bayes NoiveBayes	Weighted Avg.	0.643		0.629	0.643					
	Confusion	Matrix ===								
	a b c cla	saified as								
	2314= 10									
	2 7 b = yes									
	4						. E			

3.1.7 Classifier Algorithm Naïve Bayes

Fig 1.9 Naïve Bayes

Results:

Time taken to build the model = 0 seconds Correctly classified instances (in %) = 64.286%Total number of instances = 14

3.1.8 Classifier Algorithm Bayes Net

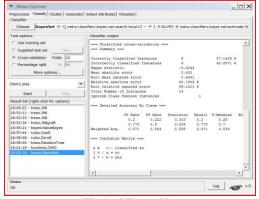


Fig.1.10 Bayes Net

Results:

Time taken to build the model = 0.01 seconds Correctly classified instances (in %) = 57.143%Total number of instances = 14



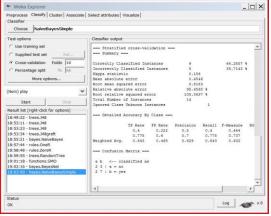


Fig 1.11 Naïve Bayes Simple

Results:

Time taken to build the model = 0 seconds

Correctly classified instances (in %) = 64.286%Total number of instances = 14

3.1.10 Classifier Algorithm Naïve Bayes Updatable

Choose NaiveBayesUpdateable									
Test options	Classifier output								
C Use training set C Supplied test set	Stratified Summary		lidation =	-					
Cross-validation Folds 10	Correctly Clas			9		64.2857			
C Percentage split % 66	Incorrectly Cl.		Instances	5		35.7143	8		
	Kappa statisti				0.186				
More options	Mean absolute Root mean squa			0.44					
in the second second	Relative absol			0.5176					
(Nom) play	Root relative			104.90					
Start Staa	Total Number o			14					
Start	Ignored Class	Unknown I	nstances		1				
Result list (right-click for options)									
18:49:22 - trees.348	Detailed A	ccuracy B	y Class						
18:53:11 - trees.J48		TP Rate	FP Bate	Precision	Recall.	F-Measure	2		
18:53:23 - trees.348		0.4	0.222	0.5	0.4	0.444	1		
18:53:34 - trees.J48graft		0.778			0.778				
18:55:21 - bayes.NaiveBayes 18:57:44 - rules.OneR	Weighted Avg.	0.643	0.465	0.629	0.643	0.632			
18:58:48 - rules.ZeroR 18:59:55 - trees.RandomTree	=== Confusion (
19:01:18 - functions.SMO	a b < cla	ssified a	5						
19:02:16 - bayes.BayesNet	231a=no								
19:03:00 - bayes.NaiveBayesSimple 19:03:43 - bayes.NaiveBayesUpdateable	2 7 i b = yes								
	4					-			

Fig. 1.12 Naïve Bayes Updatable

Results:

Time taken to build the model = 0 seconds Correctly classified instances (in %) = 64.286%Total number of instances = 14

3.1.11 Classifier Algorithm Multilayer Perceptron

Test options	Classifier output								
Use training set Supplied test set Set Folds 10	Stratified Summary	76.5714							
	Correctly Class Incorrectly Class			11		78.5714			
C Percentage split % 66	Kappa statisti		ruscances	0.51	16	21.9200			
More options	Mean absolute a			0.28					
	Root mean squar			0.48	13				
(Nom) play	Relative absolu	te error		59.39	94 8				
(Nom) pidy	Root relative			97.54	.5482 %				
Start Stop	Total Number of			14					
	Ignored Class Unknown Instances 1								
Result list (right-click for options)	1 2 2 2 2 2 2 2		12						
18:49:22 - trees.348	Detailed A	curacy B	y Class ==	-					
18:53:11 - trees.348	6	TP Bate	FP Bate	Precision	Recall.	T-Measure	RC		
18:53:23 - trees.348		0.6	0.111		0.6	0.667	-		
18:53:34 - trees.348graft		0.889		0.8	0.889	0.842			
18:55:21 - bayes.NaiveBayes 18:57:44 - rules.OneR	Weighted Avg.	0.786	0.297	0.782	0.786	0.779			
18:58:48 - rules.ZeroR	Confusion 8	atrix							
18:59:55 - trees.RandomTree									
19:01:18 - functions.SMO	a b < clas	sified a	5						
19:02:16 - bayes.BayesNet	3 2 i a = no								
19:03:00 - bayes.NaiveBayesSimple	1 8 b - yes								
19:03:43 - bayes.NaiveBayesUpdateable									
19:04:34 - functions.MultilayorPerceptron	1								
	4								

Fig. 1.13 Multilayer Perceptron

Results:

Time taken to build the model = 0.08 seconds Correctly classified instances (in %) = 78.5714%Total number of instances = 14

3.1.12 Classifier Algorithm Logistic

Classifier Choose Logistic -R 1.0E-8 -M -1							
Test options	Classifier output						
C Use training set							1923
	Stratified		idation -				-
C Supplied test set Set	=== Summary ===						
Cross-validation Folds 10	Correctly Class	ified Inc	Lances	9		64.2857	
C Percentage split % 66	Incorrectly Cla			5		35,7143	
r Percentage split % [00	Kappa statistic	3		0.18	6		
More options	Mean absolute e			0.37			
	Root mean squar			0.60			
Nom) play	Relative absolu			79.43			
	Root relative a			121.72	57 %		
Start Stop	Total Number of Ignored Class U			14	1		
Result list (right-click for options)	Ignored Class (nknown Ir	acances		*		
18:49:22 - trees.148	Detailed Ad	curacy By	Class	-			
18:53:11 - trees 148	100.00000000000000000000000000000000000						
18:53:23 - trees. J48		TP Rate	FP Rate	Precision	Recall 0.4	F-Measure	RC
18:53:34 - trees.J48graft		0.4	0.222	0.5	0.4		
18:55:21 - bayes.NaiveBayes	Weighted Avg.			0.629			
18:57:44 - rules.OneR	weighted wyg.	0.043	0.400	0.029	0.043	0.034	
18:58:48 - rules.ZeroR	=== Confusion }	atrix					
L8:59:55 - trees.RandomTree	100000000000000000000000000000000000000						
19:01:18 - functions.SMO		sified as					
19:02:16 - bayes.BayesNet	2 3 1 a = no						
19:03:00 - bayes.NaiveBayesSimple	2 7 b = yes						
19:03:43 - bayes.NaiveBayesUpdateable							-
19:04:34 - functions.MultilayerPerceptron	201						×.
19:06:36 - functions.Logistic	4						•



Results:

Time taken to build the model = 0.03 seconds Correctly classified instances (in %) = 64.286%Total number of instances = 14



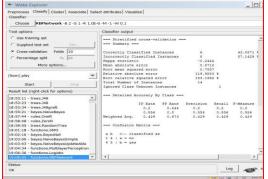


Fig. 1.15 RBF Network

Results:

♥ Wel

Time taken to build the model = 0.03 seconds Correctly classified instances (in %) = 42.857%Total number of instances = 14



Test options	Classifier output							
C Use training set C Supplied test set C Cross-validation Folds 10	<pre>=== Stratified cross-validation === === Summary === Correctly Classified Instances 8 57.14</pre>							
	Incorrectly Classified Instances	6	57.1429 %					
C Percentage split % 66	Kappa statistic	0.0667						
More options	Mean absolute error	0.4635						
And the second design of the s	Root mean squared error	0.5718						
Nom) play 🔫	Relative absolute error	97.3389 %						
	Root relative squared error Total Number of Instances	115.9073 %						
Start Stop	Total Number of Instances Ignored Class Unknown Instances	14						
Result list (right-click for options)	Agnored crass onknown inscances							
18:33:23 - trees.J46 18:33:34 - trees.J46gndt 18:35:24 - tayes.NaveBayes 18:35:44 - rules.ZeroR 19:35:44 - rules.ZeroR 19:35:45 - rules.RendomTree 19:01:16 - bayes.RendomTree 19:01:16 - bayes.Bayestightenble 19:00:16 - bayes.NaveBayestightenble 19:00:17 - bayes.NaveBayestightenble 19:00:17 - bayes.NaveBayestightenble		Precision Recall 0.4 0.4 0.667 0.667						

Fig. 1.16 Simple Logistic

Results:

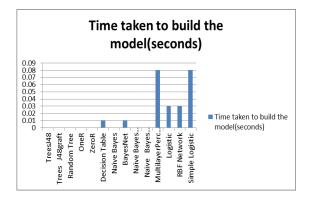
Time taken to build the model = 0.08 seconds Correctly classified instances (in %) = 57.143%Total number of instances = 14

4. CONCLUSION:

A)Table 1.3 shows the comparison of time taken to build the model with different algorithms.

	TreesJ48		Random Tree		ZeroR	Decision Table			Bayes		Multilayer Perceptron			Simple Logistic
Time taken to build the model(secon														Y
ds)	0	0	0	0	0	0.01	0	0.01	0	0	0.08	0.03	0.03	0.08

Table 1.3 Time taken (in seconds) to build the model



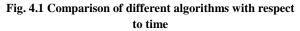


Fig. 4.1 shows that Mulitlayer Perceptron and Simple Logistic takes 0.8 seconds to build the model whereas Logistic and RBF net take 0.3 seconds each. Decision table and Bayes Net take about 0.1 seconds each. Rest of the algorithms take 0 seconds to build.

B) Table 1.4 shows the performance of classifier filters based on the identification of correct instances

		Trees			Rules			Bayes				Functions			
	TreesJ48		Random Tree			Decision Table			Bayes		Multilayer Perceptron			Simple Logisti	
Identification of Current Instances (%)	64 299	64.299	57 143	35 714	64 286	50	64 286	57 143	64.286	64 286	78.5714	64 286	42 857	57 14	

Table 1.4 Performance of classifiers based on identification of correct instances

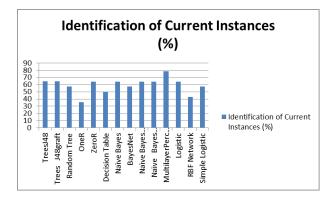


Fig 4.2 Comparison of identification of correct instances

Fig. 4.2 shows that Multilayer Perceptron has the highest identification of correct instances i.e. it is 78% for MLP and the least is that of OneR which is about 36%.

Now, if the comparison of classifier filters is done with respect to time taken to build model and identification of correct instances, then it is concluded that **Multilayer Perceptron** gives better results (time to build model = 0.08

sec, Identification of correct instances = 78.571%) compared to other classifier filters.

REFERENCES:

[1] Wikipedia.org

http://en.wikipedia.org/wiki/Weka_(machine_learning)

[2] WEKA Tutorial

http://www.cs.utexas.edu/users/ml/tutorials/Weka-tut/