Society for Science and Education
United Kingdom
TMLAI Transactions on Machine Learning and Artificial Intelligence
Volume 2, Issue 4
ISSN: 2054 - 7390

Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy

V.Mohan Patro[1], Manas Ranjan Patra[2]
Department of Computer Science, Berhampur University, Berhampur-760007, Odisha, India
[1]vmpatro@gmail.com,[2]mrpatra12@gmail.com

ABSTRACT

Accuracy of a classifier or predictor is normally estimated with the help of confusion matrix, which is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Calculation of classification accuracy of a predictor using confusion matrix for two classed attribute is simple. In case of multi classed attribute, we have to take accuracy of all the classes into consideration, to aggregate them to come with the actual accuracy of the particular classifier or predictor for that particular attribute. Here formulating this, weighted average classification accuracy has been introduced for the overall recognition rate of the classifier, which reflects how well the classifier recognizes tuples of various classes. Classification accuracy is being calculated for the classifiers BayesNet(BN), NaiveBayes(NB), J48 and Decision Table(DT) through weighted average accuracy formulation and the trend of the accuracy values for different number of instances is displayed in tables, which shows the flawless calculation.

Key words: Confusion Matrix, Classifiers, Classification Accuracy, Weighted Average Accuracy.

1 INTRODUCTION

Accuracy of a classifier on a given data set is the percentage of test set tuples that are correctly classified by the classifier. It reflects how well the classifier recognizes tuples of various classes. The error rate or misclassification rate of a classifier *M* can be expressed as 1-*Acc*(*M*), where *Acc*(*M*) is the accuracy of *M* [1].

Most common form of expressing classification accuracy is the error matrix (confusion matrix or contingency table). Error matrices compare, on a class-by-class basis, the relationship between known reference data and the corresponding results of the classification procedure.

The Overall Accuracy is computed by dividing the total number of correctly classified elements (i.e., the sum of the elements along the major diagonal) by the total number of elements in confusion matrix.

DOI: 10.14738/tmlai.24.328
Publication Date: 28th August 2014
URL: http://dx.doi.org/10.14738/tmlai.24.328

Individual Class Accuracy is calculated by dividing the number of correctly classified elements for each class by either the total number of elements in the corresponding column or row.

The Producers Accuracy is the result from dividing the number of correctly classified elements for each class (on the major diagonal) by the number of elements "known" to be of that category.

The User's Accuracy is computed by dividing the number of correctly classified elements in each class (on the major diagonal) by the total number of elements that were classified in that class.

The different types of accuracies like producer's accuracy, user's accuracy, overall accuracy etc. are being calculated with the help of different data and they are being compared [2,3,4,5,6,7,8,9]. In [2], Mittal et al. devised to compare producer and user accuracies on land cover images with the help of expectation-maximization algorithm applying on data provided by JAXA, Japan. A combination of the light detection and ranging (LiDAR) height and intensity data proved to be effective for urban land cover classification [3]. In [4], Samiappan et al. present a Non-Uniform Random Feature Selection (NU-RFS) within a Multi-Classifier System (MCS) framework and experimental results demonstrate the superiority of the proposed approach compared to SVM and RFS. In [5], Experimental results show that a multi-band and multi-level wavelet packet approach can be used to drastically increase the classification accuracy. In [6], a new method is proposed using a data structure called Peano Count Tree (P-tree) for decision tree classification and the accuracy is possessed using the parameters overall accuracy, User's accuracy and Producer's accuracy for image classification methods of object oriented classification, Knowledge Base Classification, Post classification and P-tree Classifier. In [7], a bootstrap method to quantify overall decision tree classification accuracy and confidence is described and the application of this for land use sampling strategies is discussed. Classification of waveforms is being discussed in [8]. In [9], an experiment was conducted to evaluate the differences between rule-based classifications of land cover.

2 CONFUSION MATRIX AND METHODOLOGY

A confusion matrix (*also* known as a contingency table or an error matrix) is a table layout that allows visualization of the performance of a supervised learning algorithm [10]. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. All correct guesses are located along the diagonal of the table such that errors can be easily visualized by any non-zero values outside the diagonal.

For a classifier to have good accuracy, ideally most of the tuples would be represented along the diagonal of the confusion matrix (CM). Given two classes, we can introduce the notion of positive tuples (tuples of the class, e.g., *buys computer = yes*) and negative tuples (e.g., *buys computer = no*). True positives refer to the positive tuples that were correctly labeled by the classifier, while true negatives are the negative tuples that were correctly labeled by the classifier. False positives are the negative tuples that were incorrectly labeled (e.g., tuples of class *buys computer = no* for which the classifier predicted *buys computer = yes*). Similarly, false negatives are the positive tuples that were incorrectly labeled (e.g., tuples of class *buys computer = yes* for which the classifier predicted *buys computer = no*).

2.1 Confusion Matrix for two classes

Table – 1: Confusion matrix

		Predicted Class	
		C_1	C_2
Actual Class	C_1	True positive	False negative
	C_2	False positive	True negative

C_1 – particular class C_2 – different class

True positive (TP) - The number of instances correctly classified as C_1
True negative (TN) - The number of instances correctly classified as C_2
False positive (FP) - The number of instances incorrectly classified as C_1 (actually C_2)
False negative (FN) - The number of instances incorrectly classified as C_2 (actually C_1)

P = Actual positive = TP + FN
P^1 = Predicted positive =TP + FP
N = Actual negative = FP + TN
N^1 = Predicted negative =FN + TN

TP rate = Sensitivity = TP / P = Recall
TN rate = Specificity = TN / N
FP rate = selectivity =1 – TN rate = FP / N
Precision = TP / P^1

Accuracy = (TP + TN) / (P + N)
= TP / (P + N) + TN / (P + N)
= TP / P * P / (P + N) + TN / N * N / (P + N)
= Sensitivity * P / (P + N) + Specificity * N / (P + N)

If a classification system has been trained to distinguish between cats, dogs and rabbits, a confusion matrix will summarize the results of testing the algorithm for further inspection. Assuming a sample of 27 animals — 8 cats, 6 dogs, and 13 rabbits, the resulting confusion matrix could look like the table 2.

Table - 2

		Predicted class		
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

In this confusion matrix, of the 8 actual cats, the system predicted that three were dogs, and of the six dogs, it predicted that one was rabbit and two were cats. Assuming the confusion matrix above, its corresponding table of confusion, for the cat class, would be:

Table - 3

5 true positives (actual cats that were correctly classified as cats)	3 false negatives (cats that were incorrectly marked as dogs)
2 false positives (dogs that were incorrectly labeled as cats)	17 true negatives (all the remaining animals, correctly classified as non-cats)

From above table classification accuracy for individual class cat can be obtained with the help of the formula for accuracy i.e. (TP + TN) / (P + N). The individual classification accuracy value of cat class will be (5+17) / (5+3+2+17). In this way, 2×2 matrices for dog and rabbit classes can be obtained, from which individual accuracies can be calculated.

The confusion matrix online calculator [11] gives Producer Accuracy, User Accuracy and overall accuracy. The overall accuracy is calculated as the ratio of total of diagonal elements and total elements in confusion matrix. Li Wenkai et al. discussed different formulae like evaluating classification accuracy with positive and background data in their paper [12]. The overall classifier's accuracy has been plotted for different classifiers in paper of Chitra P.K.A. et al.[13].

The overall accuracy of a classifier, in case of multi-classed attribute also, is being calculated as the ratio of total of diagonal elements and total elements in confusion matrix. It means we are taking all the true positive values of all the classes into consideration. In case of two-class attribute, true positive of one class is true negative of another class and vice-versa.

The classification accuracy is (TP + TN) / (P + N)

In our formulation for a multi-classed attribute, all the true negative values of all the classes are being taken into consideration. This means for each of the classes we put the formula of accuracy to get the individual classification accuracy of the class. Actual count of the particular class is taken as weight for the same class. Aggregating all the individual classification accuracies and weights of all the classes, the weighted average classification accuracy for the attribute is being calculated.

3 CLASSIFICATION TECHNIQUES USED

Bayesian networks are probability based and are used for the reasoning and the decision making in uncertainty, and heavily rely on Bayes' rule. Bayes' rule can be defined as follows [15];

- Assume A_i attributes where i = 1,2,3,…,n, and which take values a_i where i= 1,2,3,…,n.
- Assume C as class label and $E = (a_1, a_2,..., a_n)$ as unclassified test instance. E will be classified into class C with the maximum posterior probability. Bayes' rule for this classification is;

  $P(C \mid E) = \arg\max_{c} P(C)P(E \mid C)$

Naïve Bayesian Classifier is one of the Bayesian Classifier techniques which is also known as the state-of-the-art of the Bayesian Classifiers. In many works it has been proven that Naïve Bayesian classifiers are one of the most computationally efficient, effective and simple algorithms for Machine Learning and Data Mining applications [16]- [19]. Naïve Bayesian classifiers assume that all attributes within the same class are independent given the class label. Based on this assumption, the Bayesian rule has been modified as follows to define the Naïve Bayesian rule;

$P(C|E) = \arg\max_c P(C) \prod_{i=1}^{n} P(A_i \mid C)$

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy [20]. The training data is a set $S = s_1, s_2, \ldots$ of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i}, x_{2,i}, \ldots, x_{p,i})$ where the x_j represent attributes or features of the sample, as well as the class in which s_i falls.

Decision table is based on logical relationships just as the truth table. It is a tool that helps us to look at the combination of both completeness and inconsistency of conditions [21]. Decision tables, like decision trees or neural nets, are classification models used for prediction. They are induced by machine learning algorithms. A decision table consists of a hierarchical table in which each entry in a higher level table gets broken down by the values of a pair of additional attributes to form another table. The structure is similar to dimensional stacking.

4 WEIGHTED AVERAGE ACCURACY(WAA) ALGORITHM

Weighted average accuracy is being defined as

$\mathrm{WAA} = \frac{1}{N} \sum_{i=1}^{n} A_i * C_i$ where

n is no. of possible classes of the multi-class attribute

A_i is individual accuracy for i^{th} class

C_i is instances count for i^{th} class

N is total instances count = $\sum_{i=1}^{n} C_i$

Weighted Average Accuracy Algorithm:
Input: Confusion Matrix
Output: Weighted Average Accuracy
WAA (A, CM)
//CM is n X n confusion matrix, where n is number of classes of an attribute, on which basis //classification accuracy is calculated. A is n+1 X n+3 matrix, where first n X n is filled up with CM
BEGIN
For i=1 to n, A(n+1,i) = $\sum_{j=1}^{n} A(j,i)$ // sum of n columns
For i=1 to n+1, A(i,n+1) = $\sum_{j=1}^{n} A(i,j)$ // sum of n+1 rows, where A(n+1,n+1) is number of instances
For i=1 to n, A(i,n+2) = A(i,n+1) +A(n+1,i) – 2 x A(i,i)
// A(i,n+2) is error i.e. sum of FP & FN and A(i,i) is TP of individual class
For i=1 to n, A(i,n+3) = A(i,n+1) x [1- A(i,n+2) / A(n+1,n+1)]
// A(i,n+3) is individual weighted accuracy, where A(i,n+1) is weight
A(n+1,n+3) = $\sum_{i=1}^{n} A(i, n+3)$ / A(n+1,n+1)
// $\sum_{i=1}^{n} A(i, n+3)$ is total weighted accuracy & A(n+1,n+3) is weighted average accuracy
Return A(n+1,n+3)
END

5 EXPERIMENTATION

The techniques developed in the preceding sections have been applied on two data sets, viz., demographic data and student performance data in undergraduate examinations.

5.1 Data Set 1

This data set consists of demographic profile of citizens (UCI's census dataset) [14]. This dataset has 30162 instances with 15 attributes, such as Age, Work-class, Final-weight, Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex, Native-country, capital-gain, capital-loss, Hours-per-week, and Income. Here, the attribute on which basis the classification accuracy is to be calculated is "education". This attribute is having 16 classes i.e. Bachelors, HS-grade, 11^{th}, Masters, 9^{th}, Some-college, Assoc-acdm, 7^{th}-8^{th}, Doctorate, Assoc-voc, Prof-school, 5^{th}-6^{th}, 10^{th}, Preschool, 12^{th} & 1^{st}-4^{th}. So the classifier will give 16×16 confusion matrix.

5.2 Data Set 2

This data set involves performance of students from different backgrounds (rural/urban/distance learning/regular students) in the university examinations. The data set has 33254 instances with 10 attributes namely data-year, stream, gender, caste, rururb, gtotal, grtot, tot2, tot1, result. Classification accuracy is computed based on the attribute "result" which has 4 possible values, viz., Fail, Pass, 2^{nd} and 1^{st}. Thus, the classifier will give rise to a 4×4 confusion matrix.

5.3 WEKA Workbench

All simulations were performed in the WEKA (Waikato Environment for Knowledge Analysis) machine learning platform that provide a workbench which consists of collection of implemented popular learning schemes that can be used for practical data mining and machine learning works.

We compare the results of four classifiers BayesNet (BN), NaiveBayes (NB), J48 and Decision Table (DT). The simulations are conducted using two different test options i.e. "Use Training set" and Cross-Validation.

5.4 "Use Training set" and Cross-Validation

The "use training set" option is to train the model with whole training data. In this option, the classifier is evaluated on how well it predicts the class of the instances it was trained on. In cross-validation option, the classifier is evaluated by cross-validation, using the number of folds that are entered in the Folds text field. Here number of folds is 10. Cross-validation calculates the accuracy of the model by separating the data into two different subsets, namely, training set and validation set or testing set. The training set is used to perform the analysis and the validation set is used to validate the analysis. This testing process is continued k times to complete the k-fold cross validation procedure. We have used 10-fold cross-validation. In 10-Fold cross-validation given dataset is partitioned into 10 subsets. From these 10 subsets 9 subsets are used to perform a training fold and a single subset is used as the testing data. The process is repeated 10 times such that each subset is used as a test subset once. The estimated accuracy is then the mean of the estimates for each of the classifiers.

6 IMPLEMENTATION OF WAA ALGORITHM

For data set 1, the confusion matrix will be of 16×16 as "education" attribute is having 16 class values. The Table-4(A) displays the 16×16 confusion matrix data obtained from the classifier for the User profile dataset, having 1000 instances. This is obtained with the help of the classifier Naïve Bayes for the attribute education. It gives 16×16 matrix, because the education attribute is having 16 class values. This same data are in first 16 rows and first 16 columns of the table 5. The accuracy for the bachelors class of attribute education is calculated from the 2×2 confusion matrices i.e. given in table – 4(A & B) (in case of table-4(A), the dark lines give 2×2 matrix). This process is same as obtaining table-3 i.e. confusion matrix for the cat class from table-2 i.e. whole confusion matrix of animals.

Table-4(A)

171	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
0	325	0	0	0	4	0	0	0	0	0	0	0	0	0	0
0	0	37	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	14	0	0	1	0	0	0	0	0	0	0	0
0	2	0	0	0	216	0	0	0	3	0	0	0	0	0	0
2	0	0	0	0	0	32	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	12	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	15	0	1	0	0	0	0	0
0	0	0	0	0	2	0	0	0	46	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	11	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	9	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	17	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2

Table-4(B)

171 true positives (actual bachelors that were correctly classified as bachelors)	2 false negatives (bachelors that were incorrectly marked as Masters)
2 false positives (Assoc-acdm that were incorrectly labeled as bachelors)	825 true negatives (all the remaining education classes, correctly classified as non- bachelors)

The process of obtaining Table-4(B), 2 × 2 confusion matrix for 'bachelors' class from Table-4(A) is explained as follows. For the 1st class (bachelors), the diagonal element of row-1 & column-1 is the true positive; sum of the other elements of row-1 is false negative; sum of the other elements of column-1 is false positive; sum of rest elements is true negative. To get the confusion matrix for 2nd class (HS-grad), row-2 & column-2 are taken into consideration. This process is applicable for other classes also.

In Table-5 first 16 × 16 matrix data is actual data from the 16 × 16 confusion matrix. 17th row is meant for sum of column elements. 17th column is for sum of row elements. 18th column is for error elements i.e. sum of false positives & false negatives. 19th column is for weighted accuracy. Bold marked value is WAA.

Table-5

171	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	173	4	172.31
0	325	0	0	0	4	0	0	0	0	0	0	0	0	0	0	329	6	327.03
0	0	37	0	0	0	0	0	0	0	0	0	2	0	0	0	39	3	38.88
0	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0	55	2	54.89
0	0	0	0	14	0	0	1	0	0	0	0	0	0	0	0	15	2	14.97
0	2	0	0	0	216	0	0	0	3	0	0	0	0	0	0	221	11	218.57
2	0	0	0	0	0	32	0	0	1	0	0	0	0	0	0	35	3	34.90
0	0	0	0	1	0	0	12	0	0	0	1	0	0	0	0	14	4	13.94
0	0	0	0	0	0	0	0	15	0	1	0	0	0	0	0	16	2	15.97
0	0	0	0	0	2	0	0	0	46	0	0	0	0	0	0	48	6	47.71
0	0	0	0	0	0	0	0	1	0	11	0	0	0	0	0	12	2	11.98
0	0	0	0	0	0	0	1	0	0	0	9	0	0	0	0	10	4	9.96
0	0	1	0	0	0	0	0	0	0	0	0	17	0	0	0	18	3	17.95
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2.00
0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	9	0	9.00
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	4	4	3.98
173	327	38	57	15	222	32	14	16	50	12	12	19	0	9	4	1000		**0.994027**

Formulation of WAA from Confusion Matrix for multi-class attribute :-

Confusion Matrix – n x n → A(n,n)

1	2 …	n	n+1	n+2	n+3
A(1,1)	A(1,2)…..	A(1,n)	$\sum_{i=1}^{n} \mathrm{A}(1,i)$		
A(2,1)	A(2,2)……	A(2,n)	$\sum_{i=1}^{n} \mathrm{A}(2,i)$		
.					
.					
.					
A(n,1)	A(n,2)…..	A(n,n)	$\sum_{i=1}^{n} \mathrm{A}(n,i)$		
$\sum_{i=1}^{n} \mathrm{A}(i,1)$	$\sum_{i=1}^{n} \mathrm{A}(i,2)$….	$\sum_{i=1}^{n} \mathrm{A}(i,n)$			

1. For i=1 to n, A(i,n+1) = $\sum_{j=1}^{n} \mathrm{A}(i,j)$ → for n+1 column
2. For i=1 to n, A(n+1,i) = $\sum_{j=1}^{n} \mathrm{A}(j,i)$ → for n+1 row
3. For i=1 to n, A(i,n+2) = A(i,n+1) +A(n+1,i) – 2 x A(i,i) → for n+2 column
4. For i=1 to n, A(i,n+3) = A(i,n+1) x [1- A(i,n+2) / A(n+1,n+1)] → for n+3 column, where A(n+1,n+1) is total number of instances.
5. Weighted Accuracy = $\sum_{i=1}^{n} \mathrm{A}(i,n+3)$
6. **Weighted Average Accuracy,** A(n+1,n+3) = $\sum_{i=1}^{n} \mathrm{A}(i,n+3)$ / A(n+1,n+1)

6.1 Data Set 1

The overall accuracy (OA) i.e. the sum of diagonal elements / the sum of all elements and WAA are being calculated for different classifiers with various numbers of instances. From table-6(A) it is clear that WAA is giving high precision value in comparison to OA.

Table-6(A)

Classifiers	1000 (OA)	1000 (WAA)	30162(OA)	30162 (WAA)
BN	0.959	0.993829	0.99817651	0.99986132
NB	0.971	0.994027	0.99472847	0.99927758
J48	0.998	0.999988	1.00000000	1
DT	0.998	0.999650	1.00000000	1

In Table-6(B) the values for weighted average accuracies as well as overall accuracies are given.

These values are result of the simulations of the classifier Bayes Net for instances 1000, 5000, 10000, 15000, 20000 and 30162. It is clear that WAA values are having high precision and consistency.

Table-6(B)-1

	1000	5000	10000
OA	0.959	0.992	0.9955
WAA	0.993829	0.9961204	0.99967512

Table-6(B)-2

	15000	20000	30162
OA	0.997	0.99765	0.99817651
WAA	0.99973017	0.99980905	0.99986132

The weighted average accuracy is calculated for different number of instances i.e. 1000, 5000, 10000, 15000, 20000, 25000 & 30162 for four classifiers i.e. Bayes Net (BN), Naïve Bayes (NB), J48 and Decision Table (DT) for the test option cross-validation. These values are given in the table 7.

Table-7(A)

	1000	5000	10000	15000
BN	0.993829	0.9961204	0.99967512	0.99973017
NB	0.994027	0.9986268	0.9990493	0.99920068
J48	0.999988	1	1	1
DT	0.999650	1	1	1

Table-7(B)

	20000	25000	30162
BN	0.99980905	0.99983766	0.99986132
NB	0.99920948	0.99926319	0.99927758
J48	1	1	1
DT	1	1	1

In fig-1 the graphs have been plotted for the above values, which show the increasing order of the accuracy is so clear.

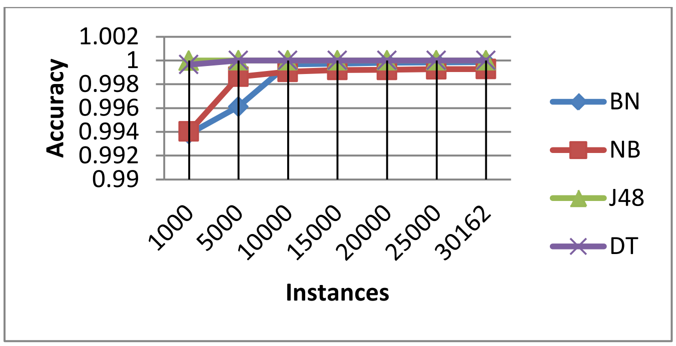

Fig-1

The weighted average accuracy is calculated for different number of instances i.e. 5000, 10000, 15000 & 20000 for the classifier Bayes Net (BN) for the test options use training set (UTS) & cross-validation (CV). The values are given in table 8.

Table-8

BN	5000	10000	15000	20000
UTS	0.99916259	0.99975627	0.99983151	0.99986265
CV	0.9961204	0.99967512	0.99973017	0.99980905

In fig-2 the graphs have been plotted for the above values. In different options also it shows the increasing order of the accuracy is clear.

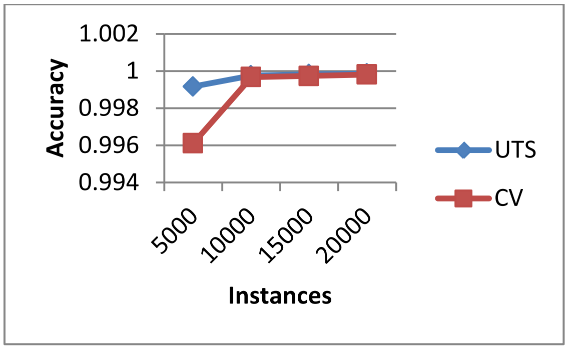

Fig-2

The weighted average accuracy is calculated for different number of instances i.e. 5000, 10000, 15000 & 20000 for the classifiers Naïve Bayes (NB) for the test options use training set & cross-validation . The values are in Table-9 and graphs are in fig-3.

Table-9

NB	5000	10000	15000	20000
UTS	0.99910824	0.99927524	0.999394947	0.999400623
CV	0.9986268	0.9990493	0.999200676	0.999209475

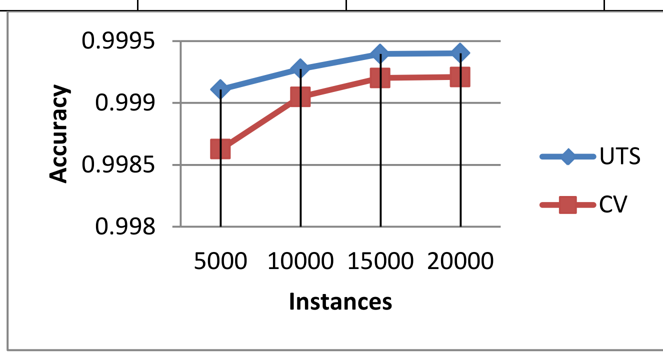

Fig-3

The above process is repeated for the classifier J48 and accuracy values are in table-10 and graphs are in fig-4.

Table - 10

J48	5000	10000	15000	20000
Use training set	1	1	1	1
Cross-validation	1	1	1	1

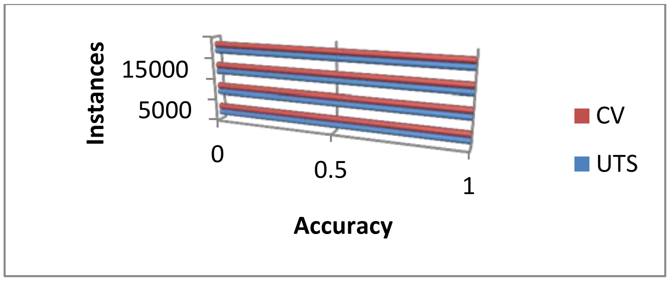

Fig-4

6.2 Data Set 2

In order to verify the accuracy of our proposed algorithm, we have also applied the technique on another data set. Table-11(A), 11(B)-1, and 11(B)-2 summarize the results obtained on the second data set. It is observed that the classifiers behave in a similar fashion confirming that the accuracy increases to a considerable extent as the number of instances grows. Thus, one can establish the supremacy of "Weighted Average Accuracy" technique over the "Overall Accuracy" computation irrespective of the data set used.

Table-11(A)

Classifiers	1000 (OA)	1000 (WAA)	33254 (OA)	33254(WAA)
BN	0.787	0.890326	0.868106	0.92404887
NB	0.794	0.894396	0.855266	0.915228077
J48	0.783	0.88757	0.880014	0.932542206
DT	0.779	0.885846	0.866633	0.923259574

Table-11(B)-1

	5000	10000	15000
OA	0.7938	0.8276	0.844333
WAA	0.89446528	0.90468752	0.912737067

Table-11(B)-2

	20000	25000	30000
OA	0.85485	0.8602	0.863867
WAA	0.917283103	0.919080608	0.921331164

7 CONCLUSION

The aim of our work was to enhance the accuracy of any classifier. Towards this end, we have formulated a technique called weighted average accuracy which is obtained by aggregating the individual accuracies for all class values of the particular attribute using the weight factor. The WAA takes the number of particular class in an attribute as the weight factor to calculate the classification accuracy. Individual accuracy is calculated with this weight factor. Lastly average of the total weighted accuracy is taken as final value. From both the data sets it is observed that for any number of instances, for any classifier, WAA out performs OA.

REFERENCES

[1] Jiawei Han and Micheline Kamber, *Book on"* Data Mining: Concepts and Techniques*", 2nd ed.,* Morgan Kaufmann Publishers, March 2006. ISBN 978-1-55860-901-3.

[2] Vikas Mittal, D. Singh and L.M. Saini; "Land Cover Classification using EM Algorithm based Multi-Polarized ALOS PALSAR Image Fusion"; IEEE, 2013, Page(s) 5pp.

[3] Weiqi Zhou; "An Object-Based Approach for Urban Land Cover Classification Integrating LiDAR Height and Intensity Data"; IEEE, 2013, pp.928-931.

[4] Sathish kumar Samiappan, Saurabh Prasad and Lori M Bruce; "Non-Uniform Random Feature Selection and Kernel Density Scoring With SVM Based Ensemble Classification for Hyperspectral Image Analysis"; IEEE, 2013, pp. 792-800.

[5] S. Rajesh and S. Arivazhagan ; "Land CoverLand Use Mapping using Different Wavelet Packet Transforms for LISS IV Imagery"; IEEE, 2011, pp. 103-108.

[6] M. Seetha, K. V. N. Sunitha, D.V. Lalitha Parameswari, G. Ravi; "Accuracy Assessment of Object Oriented and Knowledge Base Image Classification using P-Trees"; IEEE, 2010, pp. 760-763.

[7] Catherine Champagne, Heather McNairn, Bahram Daneshfar, Jiali Shang; "A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada"; Elsevier, 2014, pp. 44-52.

[8] Karolina D. Fieber, Ian J. Davenport, James M. Ferryman, Robert J. Gurney, Jeffrey P. Walker, Jorg M. Hacker, "Analysis of full-waveform LiDAR data for classification of an orange orchard scene"; Elsevier, 2013, pp. 63-82.

[9] Mariana Belgiu, Lucian Draˇgut, Josef Strobl; "Quantitative evaluation of variations in rule-based classifications of land cover in urban neighborhoods using WorldView-2 imagery"; Elsevier, 2013, pp. 205-215.

[10] http://en.wikipedia.org/wiki/Confusion_matrix last accessed on 10/04/13.

[11] http://www.dicom.uninsubria.it/~marco.vanetti/cfmatrix last accessed on 15/05/13.

[12] Wenkai Li and Qinghua Guo, "A New Accuracy Assessment Method for One-Class Remote Sensing Classification", IEEE, 2013, pp. 1-12.

[13] P.K.A. Chitra and S. Appavu Alias Balamurugan, "Benchmark Evaluation of classification methods for single label learning with *R",* IEEE, 2013, pp. 746-752.

[14] Asuncion A. and Newman D.J. (2007) UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. [Online] Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html last accessed on 11/02/13.

[15] Jensen F.V. "Introduction to Bayesian Networks". Denmark: Hugin Expert A/S, 1993.

[16] Wang Z. and I. Webb G., "Comparison of lazy bayesian rule and tree-augmented bayesian learning", IEEE, 2002, pp. 490 – 497.

[17] Shi Z., Huang Y. and Zhang S., "Fisher score based naive Bayesian classifier", IEEE, 2005, pp. 1616-1621.

[18] Xie Z. and Zhang Q., "A study of selective neighborhood-based naïve bayes for efficient lazy learning". 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2004.

[19] Santafe G., Loranzo J.A. and Larranaga P., "Bayesian model averaging of naive bayes for clustering", IEEE, 2006, Page(s) 1149 – 1161.

[20] http://en.wikipedia.org/wiki/C4.5_algorithm last accessed on 21/03/13.

[21] Alicia Y.C. Tang, Nur Hanani Azami and Norfaezah Osman, "Application of Data Mining Techniques in Customer relationship Management for An Automobile Company", IEEE, 2011, Page(s) 7pp.

