Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 5, May - 2015

Seed Classification using Machine Learning Techniques

Raja Haroon Ajaz
Department of Computing and IT
Iqra University, Islamabad Campus, Pakistan
Email: haroon11_raja@yahoo.com

Lal Hussain
Quality Enhancement Cell/ DCS &IT,University of Azad Jammu and Kashmir
Muzaffarabad, Pakistan
Email: lall_hussain2008@live.com

***Abstract*—Seed classification is a process in which different varieties of seeds are categorized into different classes on the basis of their morphological features. In the present work We performed seed classification using Weka tool. The data was collected from UCI website's database. The features of seed used are area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove. In Weka classification we used Function, Bayes, Meta and Lazy methods. The classifiers used from these methods are Multilayer Perceptron, Logistics, SMO, NaiveBayes Updateable, Naïve Bayes, Bayes Net, MultiClass Classifier, Classification Via Regression and LWL. After that we used 10 fold, 5 fold and 2 fold Cross Validation as well as Training Set method. Multilayer Perceptron and Logistics from function method gives higher accuracy than all other classifiers, which is 95.2% for both the classifiers using 10 fold Cross Validation. We also analyzed the results using 5 fold and 2 fold Cross Validation, but we observed that the overall performance measures decreases as we decrease the fold value except the Multilayer Perceptron classifier that gives the highest accuracy value 97.6% using 5 fold Cross Validation. Multilayer Perceptron also gives the highest accuracy value when we use Training Set method which is 99.5% and Logistics gives second highest accuracy value of 98.6%. Finally we also observed that Training Set method gives higher accuracy than Cross Validation during the classification process.**

***Keywords*—Classification, Classifier, Cross Validation, Receiver Operator Curve (ROC)**

I. INTRODUCTION

A comparison of a supervised (Back Propagation) and an un-supervised (Self Organizing Map) artificial neural network is made to classify the chickpea seed varieties in this paper. This research is concluded as the unsupervised artificial neural network gives better performance with 79% accuracy as compared to the supervised artificial neural networks which gives 73% accuracy. The classification of chickpea seeds varieties was made according to the morphological properties of chickpea seeds, by considering its 400 samples which includes its four varieties; Kaka, Piroz, Ilc and Jam [1].

The ANN modeling becoming very popular in different areas of agriculture, specially, in the areas where straight statistical modeling becomes unsuccessful. The ANN is using in the field of agriculture to predict the crop yield, biomass production, seeding dates, physical and physiological damaging of seeds, organic matter contents in the soils, soil moisture estimation, aerodynamic properties of crops, estimation of sugar content in fruits and characterization of crop varieties [2].

In this paper shows the capability and potential of machine vision with the well- trained multilayer neural network classifiers for shapes, sizes, and varietal type identification of irregular rice grain samples grown in the assorted agro environmental zones in the country. A machine vision composed with the established neural network architectures could be used as a tool to attain better and more impartial rice quality evaluation according to the business point of view [3].

The machine vision gained very much importance in agricultural industry. Classification and seed analysis can deliver additional information and knowledge in their production, seeds quality control and its contaminations identification. Normally these tasks and activities are accomplished by specialists by visually reviewing each sample, which is a very wearisome and time consuming activity. Computer Vision technology is applied for the inspection of quality of corn seed for achieving the accurate and fast inspection performance. To attain the identical quality standard from various inspection staffs those have various levels of experience and skill is a hard hitting task. This method is suggested for the classification of the quality of corn seeds according to their defects types [4].

Machine vision techniques are practical in a large array of fields to improve efficiency of the specified types of work. In the paper machine vision technique is used for the recognition characteristic of the classification problems [5].

Seed inspection and grading in the classification of different varieties of seeds are generally determined with respect to morphological features and color of seeds because of the the main visual factors in seed classification process. In certain crops as corn, because of dissimilarities between varietal morphology and quality, the seeds identification is very significant. [6].

Pazoki and Pazoki (2011) to classify 5 varieties of rain fed wheat grain cultivar artificial neural network is used and the average accuracy that gained was 86.48%. But after feature selection when UTA algorithm is used the accuracy increased to 87.22% [7].

Chen et al 2010, suggested a model as a combination of vision-based approach and pattern recognition techniques along with the neural networks to classify the five corn seed varieties. The results exhibited average classification accuracy up and around 90% for five stated varieties of the corn seeds [8].

A machine vision system is an alternate to the manual inspection, in the field of biological sciences to analyze biological products. To classify the varieties of various food crops and for identifying their quality as well, the machine vision is broadly used in the field of agriculture. Machine algorithms can be used to identify different varieties of wheat seeds to classify them according to their quality [9].

It's getting importance to transfer the technology for identification of the quality of seeds. Hence, machine learning technique works to classify seeds quality on the basis of different stages of the cotton crop growth. In machine learning approach following classifiers are used such as Multilayer Perceptron, Decision Tree and Naive Bayes Classifier in the specified model. The supply and production of the cotton seed of enriched varieties varies from the evaluation of the varieties and is a very technical and critical task. In the specified work machine learning techniques are used to concentrate on the several growth periods of the crop cotton and classification of seed cotton [10].

In biology and agronomy crop seed characteristics are very significant aspects. Machine vision technology is developed to quantify the features, the quality precise examination and graduation of the crop seeds. A novel scheme is presented to extract and quantify some of features having worth biologically. A system regulation method with VCD disc device has used as the point of orientation for extracting equivalent diameter of seed. To qualify the plumpness of rapeseed seed, variation coefficient of the radius has applied. Also the major color means and nine colors HSV model has applied to qualify and identify the rapeseed seed color [11].

II. DATA ACQUISITION

The data of wheat seeds is gathered from **UCI** website which is a great dataset repository. The numbers of samples of wheat seeds are 210 from three wheat classes Kama, Rosa and Canadian are collected for classification process. Seven geometrical or morphological features of seeds are considered on the basis of which seeds are classified into three classes of wheat.

III. PROPOSED METHODOLOGIES

Classification through Weka software is used here for the analysis of structural activity relationship. The above data was then processed for classification which was prepared in ARFF format. Four classification methods are used in our system such as Function, Bayes, Meta and Lazy on seeds datasets to test the classification performance in this system.

3.1.1 Bayes Method:

In Bayes method Bayes Net (BN), Bayesian Logistic Regression (BLR), Complement Naive Bayes (CNB), Naive Bayes (NB), DMNB Text, Naive Bayes Multinomial (NBMN), Naive Bayes Simple (NBS), Naive Bayes Multinomial Updateable (NBMNU), and Naive Bayes Updateable (NBU) are included.

3.1.1.1 Naive Bayes classifier:

Naive Bayes classifier is a set of supervised learning algorithms based on application of Bayes theorem with strong assumptions of independence between every pair of features and is a simple probabilistic classifier. In a supervised learning setting training of Naive Bayes classifiers can efficiently done, and it depends on the probability model's precise nature.

3.1.1.2 Logistics:

Logistic classifier is used for measuring the association between one or more independent variables and a clear-cut dependent variable, usually these are continuous.

3.1.2 Functions Method:

Function Method includes Multilayer Perceptron (MP), Logistic, SMO, Radial Base Function Network (RBFN), LibLinear (LL), LibSVM (LSVM), SPegasos (SP), Voted Perceptron (VP) and Simple Logistic (SL).

3.1.2.1 SMO:

Sequential Minimal Optimization (SMO) algorithm is invented by John Platt in 1998. SMO algorithm is widely used to solve the optimization problem and in the Support Vector Machine's training.

3.1.2.2 Multilayer Perceptron:

Multilayer perceptron is basically a feed forward ANN model. MLPs are the universal approximators that are used to map sets of input data onto a set of suitable output. It gets training with one or more hidden layers by using Weka's optimization class by minimizing the squared error. Specific weight wij is allocated to every node in one layer. Several parameters are there. Ridge is a parameter that determines the penalty on the size of the weights. Number of hidden units can also be indicated. Large numbers increase training times, and conjugate gradient descent can also be used in place of BFGS updates, which is somewhat faster for cases with many parameters. Unsupervised Nominal to Binary filter is used to process the nominal attributes and through Replace Missing Values, missing values are replaced globally.

3.1.3 Lazy Method:

Lazy method includes LWL, Kstar, IBK and IBI.

3.1.3.1 Lazy LWL: Lazy LWL abbreviated as Locally Weighted Learning algorithm is usually involved in finding the relevant and to store the training data into the memory. Locally weighted learning algorithm is used to transfers weights to the instances, hence called instance or case algorithm.

3.1.4 Meta Method:

Meta method includes MultiClass Classifier (MCC), Classification via Regression (CR), AdaBoostMI (ABMI), Classification via Clustering (CC), Bagging, Decorate, Grading, Dagging, LogiBoost (LB), Filtered Classifiers and MultiBoost AB (MBAB).

3.2 Cross Validation

Cross-Validation is used to compare and evaluate the learning algorithms. It divides the data into two sectors, one of them is used for learning or training a model, but the other is used for validation of the model, so it is a statistical method. Normally in cross-validation, validation sets and the training set must cross over in successive rounds, though there should be a chance of validation against each data point. There are many forms of cross validation but k-fold cross validation is the basic form of cross-validation method. Cross validation method have some more forms like repeated rounds of k-fold cross-validation and special cases of k-fold cross-validation.

3.2.1 K-Fold Cross-Validation

K-fold cross validation make the partitions of data into k equal or nearly equal sized folds or segments. Then k iterations of validation and training are performed. And within each iteration different segments of the data is detained out for validation. The remaining k-1 folds are used for learning. To being data split into k folds, data is usually stratified earlier. In this process a good representation of data as a whole is confirmed and rearrangement is done if needed. Binary classification problem is an example where each class comprises half of the data. It is great to organize the data such that in every fold, around half of the instances should be in each class.

IV. RESULTS AND DISCUSSIONS

A receiver operating characteristics graphs are used to visualize, select and organize all the classifiers on the basis of their performance. ROC graphs (Swets et al., 2000 and Egan, 1975) are in practice since long time for detecting signals and showing a tradeoff between the classifier's false alarm rate and the hit rate.

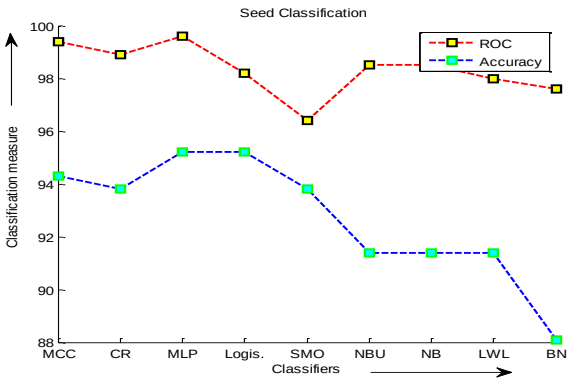

Figure 1: Seed Classification Using 10 Fold Cross Validation

Multiclass Classifier (MCC), Classification via Regression(CR), Multilayer Perceptron(MLP), Logistics (Logis.), SMO, NaiveBayesUpdateable(NBU), NaiveBayes(NB), LWL, BayesNet(BN)

The above Figure 1 and Table 10 shows the seed classification using 10 fold cross validation. The classifiers used are Multiclass Classifier (MCC), Classification via Regression (CR), Multilayer Perceptron (MLP), Logistics (Logis.), SMO, NaiveBayesUpdateable (NBU), NaiveBayes (NB), LWL and BayesNet (BN). The classifiers chosen from Weka classification methods those have highest accuracy among all the others. The features used for classification are area, compactness, perimeter, width of kernel, length of kernel, length of kernel groove and asymmetric coefficient. Seventy samples of each seed such as Kama, Rosa and Canadian are taken. The figure 1 depicts the accuracy and ROC values against each of the classifier. From the figure it is seen that the classifier Multilayer Perceptron and Logistic gives the highest accuracy that is 95.2 % among all other classifiers when used 10-Fold Cross Validation. Second highest accuracy obtained from Multi Class Classifier which is 94.3 %, while the Bayes Net classifier gives lower classification accuracy of 88.1 %. Similarly, the highest ROC values obtained from Multilayer Perceptron which is 99.6% and second highest from Multiclass Classifier which is 99.4%.

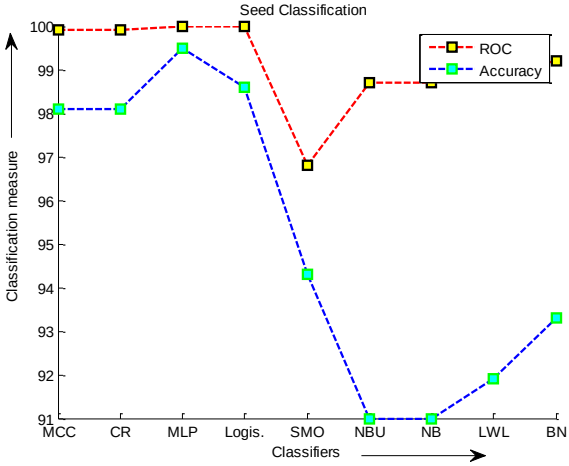

Figure 2: Seed Classification Using 5 Fold Cross Validation

Multiclass Classifier (MCC), Classification via Regression(CR), Multilayer Perceptron(MLP), Logistics (Logis.), SMO, NaiveBayesUpdateable(NBU), NaiveBayes(NB), LWL, BayesNet(BN)

The above Figure 2 and Table 2 show the seed classification using 5-Fold Cross Validation. The figure 2 shows the accuracy and ROC values obtained using the above mentioned classifiers. It is seen from the Figure 2 and Table 2 that again the Multilayer Perceptron gives the highest accuracy of 97.6 % and Bayes Net gives an accuracy of 87.6%. However, the accuracy values slightly decreases against each classifier while ROC values increases in some of classifier as depicted in the Figure 2 and Table 2. Hence, we can conclude that the overall accuracy decreases as we decrease the cross validation fold.

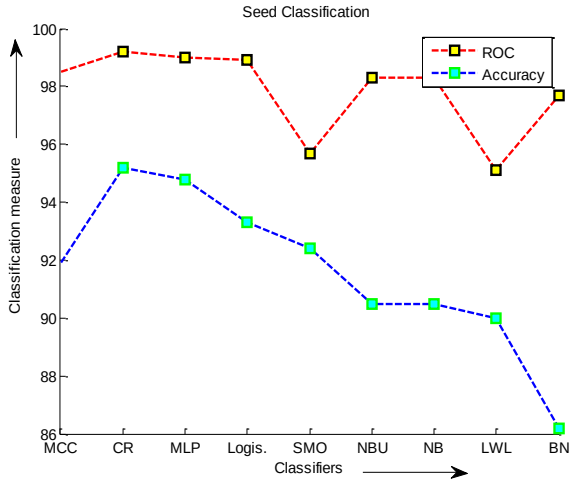

Figure 3: Seed Classification Using 2 fold Cross Validation

The above Figure 3 shows the seed classification using 2-Fold Cross Validation. Using 2-Fold Cross validation, we observe that accuracy of all the classifiers again decreases slightly except the Classification via Regression classifiers which is 95.2%, 94.3%, 93.8 % using 2-Ffold CV, 5-Fold CV and 10- Fold CV respectively.

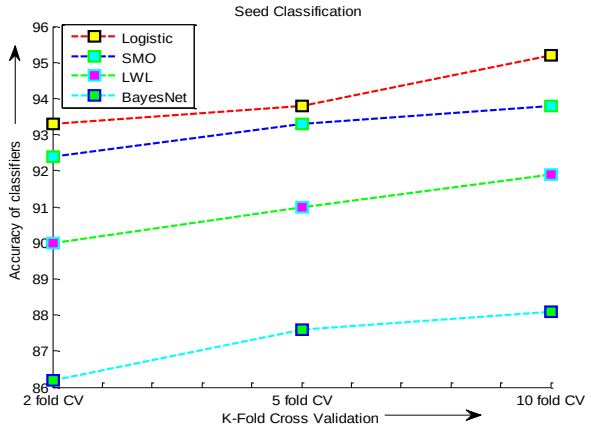

Figure 4: Accuracy of different classifiers using 2, 5 and 10 Fold CV

The above Figure 4 depicts the classifiers accuracy using 2-Fold, 5-Fold and 10-Fold Cross validation. The classifiers used here are Logistic, SMO, LWL and Bayes Net. In all these classifiers the accuracy increases from fold 2 to fold 10 cross validation except MLP. Thus, we can conclude that 10-Fold Cross validation gives the overall best classification rate in this situation. By decreasing the k value, the accuracy measure of most of the classifiers also decreases.

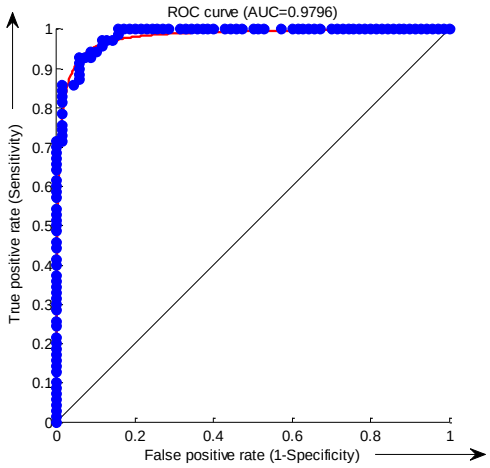

Figure 5: ROC Curve Analysis of Kama and Rosa Seeds using Perimeter as its features

In the Figure 5 above, the ROC AUC obtained is 0.9796 to classify the Kama and Rosa seed using perimeter as its feature. The AUC value shows that perimeter feature is best to classify the Kama and Rosa seed.

The Tables 1, 2 and 3 below also depicts the seed classification using k-fold CV and training set methods. From these Tables it is clearly depicted that Training set methods gives higher TP rates than k-fold CV methods. As the higher TP rate was obtained using Multilayer Perceptron (TP rate=99.5%) using training set method, while (TP rate=97.6%) using 5-fold CV method and (TP rate=95.2%) using 10 fold CV method. The other classifiers also represent the similar behavior and accuracy level as shown below.

TABLE 1: SEED CLASSIFICATION USING 10 FOLD CROSS VALIDATION

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
MultiClass Classifier	94.3%	02.9%	94.3%	94.3%	94.3%	99.4%
Classification ViaRegression	93.8%	03.1%	93.8%	93.8%	93.8%	98.9%
Multilayer Perceptron	95.2%	02.4%	95.3%	95.2%	95.2%	99.6%
Logistics	95.2%	02.4%	95.3%	95.2%	95.2%	98.2%
SMO	93.8%	03.1%	93.8%	93.8%	93.8%	96.4%
NaiveBayesUpdateable	91.4%	04.3%	91.4%	91.4%	91.4%	98.5%

NaiveBayes	91.4%	04.3%	91.4%	91.4%	91.4%	98.5%
LWL	91.9%	04.0%	92.8%	91.9%	91.8%	98.0%
BayesNet	88.1%	06.0%	88.2%	88.1%	88.0%	97.6%

TABLE 2: SEED CLASSIFICATION USING 5 FOLD CROSS VALIDATION

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
MultiClass Classifier	95.7%	02.1%	95.8%	95.7%	95.7%	99.4%
Classification ViaRegression	94.3%	02.9%	94.3%	94.3%	94.2%	98.9%
Multilayer Perceptron	97.6%	01.2%	97.6%	97.6%	97.6%	99.8%
Logistics	93.8%	03.1%	93.8%	93.8%	93.8%	97.1%
SMO	93.3%	03.3%	93.3%	93.3%	93.3%	96.2%
NaiveBayesUpdateable	91.0%	04.5%	91.0%	91.0%	90.9%	98.5%
NaiveBayes	91.0%	04.5%	91.0%	91.0%	90.9%	98.5%
LWL	91.0%	04.5%	91.6%	91.0%	90.9%	96.8%
BayesNet	87.6%	06.2%	87.5%	87.6%	87.5%	97.5%

TABLE 3: SEED CLASSIFICATION USING TRAINING SET METHOD

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
MultiClass Classifier	98.1%	01.0%	98.1%	98.1%	98.1%	99.9%
Classification ViaRegression	98.1%	01.0%	98.1%	98.1%	98.1%	99.9%
Multilayer Perceptron	99.5%	00.2%	99.5%	99.5%	99.5%	100%
Logistics	98.6%	00.7%	98.6%	98.6%	98.6%	100%
SMO	94.3%	02.9%	94.3%	94.3%	94.3%	96.8%
NaiveBayesUpdateable	91.0%	04.5%	91.0%	91.0%	90.9%	98.7%
NaiveBayes	91.0%	04.5%	91.0%	91.0%	90.9%	98.7%
LWL	91.9%	04.0%	92.8%	91.9%	91.8%	98.9%
BayesNet	93.3%	03.3%	93.7%	93.3%	93.3%	99.2%

V. CONCLUSION

In the present work we have classified the, Kama, Rosa and Canadian seeds using Weka classification algorithms and methods. The data set was taken from UCI website's database. The performance was measured using K-Fold cross validations and Training set method. Multilayer Perceptron MLP using 5-Fold cross validation gives highest performance of 97.6% among all the Weka classifier whereas MLP also gives highest performance of 99.5% among all other Weka classifiers when we use training set method. We noted that training set gives higher performance than cross validation method. We also observed that the performance decreases as the number of folds decreases and all the classifiers except the MLP that gives highest performance on 5-Fold cross validation.

In future, we will use some other machine learning classifiers other than the Weka classifiers. We will also combine the classifiers to further improve the performance using Bagging and Boosting techniques. Besides there are also un-supervised machine learning techniques, such as clustering to classify the seeds. Moreover we can classify the other categories of seeds using these techniques and classifiers after extracting these features.

REFERENCES

[1]Ghamari, S. (2012). Classification of chickpea seeds using supervised and unsupervised artificial neural networks. African Journal of Agricultural Research, 7(21), 3193-3201.

[2]Ghamari, S., Borghei, A. M., Rabbani, H., Khazaei, J., & Basati, F. (2010). Modeling the terminal velocity of agricultural seeds with artificial neural networks. Afr. J. Agric. Res, 5(5), 389-398.

[3]Guzman, J. D., & Peralta, E. K. (2008). Classification of Philippine Rice Grains Using Machine Vision and Artificial Neural Networks. In World conference on agricultural information and IT.

[4]Kantip Kiratiratanapruk and Wasin Sinthupinyo.(2011), Color And Texture For Corn Seed Classification By Machine Vision, International Symposium on Intelligent Signal Processing & Communication Systems (ISPACS). pp 1-5.

[5]Adjemout, O., Hammouche, K., & Diaf, M. (2007, February). Automatic seeds recognition by size, form and texture features. In Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on (pp. 1-4). IEEE.

[6]Zapotoczny, P., Zielinska, M., & Nita, Z. (2008). Application of image analysis for the varietal classification of barley:: Morphological features. Journal of Cereal Science, 48(1), 104-110.

[7]Pazoki, A., & Pazoki, Z. (2011). Classification system for rain fed wheat grain cultivars using artificial neural network. African Journal of Biotechnology, 10(41), 8031-8038.

[8]Chen, X., Xun, Y., Li, W., & Zhang, J. (2010). Combining discriminant analysis and neural networks for corn variety identification. Computers and electronics in agriculture, 71, S48-S53.

[9]Punn, M., & Bhalla, N. (2013). Classification of Wheat Grains Using Machine Algorithms. International Journal. (IJSR), 2319-7064.

[10]KS, J. (2010, June). Classification of Seed Cotton Yield Based on the Growth Stages of Cotton Crop Using Machine Learning Techniques. In Advances in Computer Engineering (ACE), 2010 International Conference on (pp. 312-315). IEEE.

[11]Li, J., Liao, G., Ou, Z., & Jin, J. (2007, December). Rapeseed seeds classification by machine vision. In Intelligent Information Technology Application, Workshop on (pp. 222-226). IEEE.

[12]M. Costa et al., Phys. Rev. E 71 (2005) 021906