International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
International Conference on Technological Emerging Challenges (ICTEC-2019)
Available online at www.ijrat.org

A COMPARISON OF VARIOUS MACHINE LEARNING ALGORITHMS FOR WHEAT SEED DATASET CLASSIFICATION

B. Gnana Priya
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

Abstract: Machine Learning is the fast emerging technology which is poised to dominate in almost all walks of life. Machine Learning not only analyses data, but also predicts future responses/actions aimed towards greater results. Artificial Intelligence will be the prime support system for human resources, human actions and performances in all areas of application. In this work different variety of wheat Seeds were classified. The category of the seed is decided based on their morphological features. Algorithms such as K-nearest neigbour, Support vector machine, Decision Trees, Random Forest and Naive Bayes classifier were applied to the Wheat seed dataset and were compared.

KEYWORDS: Machine learning, KNN, SVM, Naive Bayes, Decision Trees, Random Forest

1. INTRODUCTION

Machine learning is employed in almost every field nowadays. Starting from the recommendations based on our interest to filtering systems in our email inbox everyone is somewhere dependent on one or several of the machine learning systems. With internet becoming more personalised, machine learning is very popular and is the key component of the future. Machine learning is making the system to act without being explicitly programmed that is allowing the computer to learn automatically. The various applications includes video surveillance, e-mail spam filtering, online fraud detection, virtual personal assistance like Alexa, automatic traffic prediction using GPS and many more.

Machine learning algorithms are broadly classified into supervised and unsupervised algorithms. Supervised learning is a method in which we train the machine using data which are well labelled or the problem for which answers are well known. Then the machine is provided with new set of examples and the supervised learning algorithm analyses the training data and comes out with predictions and produces an correct outcome from labelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Reinforcement learning approach is based on observation. The network makes a decision by observing the environment. If observation is negative, the network adjusts its weights to be able to make a different required decision the next time.

2. RELATED WORKS

Ahamed [2] developed an automated system to analyse and classify wheat seeds using k-means clustering algorithm. A higher confidence level and faster classification rate is achieved. Haroon [3] performs seed classification using Weka tool. Multifold cross validation is employed. L. Lin and et al [4] introduced a method based on fuzzy theory by considering the characteristics of wheat seed which helps in recognition the seed type. Tabu search method employed. M. R. Neuman and et al [5] developed a workstation assisting in cereal grain inspection for classifying purposes video colorimetry methodology is proposed to help measuring color of cereal grains. 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 [6]. 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 [8].

3. PROPOSED WORK

3.1. Wheat Seeds Dataset

The Wheat Seeds Dataset involves the prediction of species given measurements of seeds from different varieties of wheat. The number of observations for each class is balanced. The examined group comprises of kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. The variable names are as follows:

1. Area.
2. Perimeter.
3. Compactness
4. Length of kernel.
5. Width of kernel.
6. Asymmetry coefficient.
7. Length of kernel groove.
8. Class (1, 2, 3).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 28%. The data was collected from UCI website's database.

Area	Perimeter	Compactness	Length of kernel	Width of kernel	Asymmetry coefficient	Length of kernel groove	Class
15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
16.82	15.51	0.8786	6.017	3.486	4.004	5.841	2
16.77	15.62	0.8638	5.927	3.438	4.92	5.795	2
17.32	15.91	0.8599	6.064	3.403	3.824	5.922	2
10.74	12.73	0.8329	5.145	2.642	4.702	4.963	3
11.48	13.05	0.8473	5.18	2.758	5.876	5.002	3
12.21	13.47	0.8453	5.357	2.893	1.661	5.178	3

Table 1: Sample from WheatSeed Dataset

3.2 Scikit Learn

Scikit-learn is a machine learning library for python. It consists of various inbuilt algorithm like support vector machine, Random forest, Decision trees, K-nearest neighbour and more. We use the train_test_split method to split the dataset into training and testing samples. In order to make predictions we use KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, svm and GaussianNB. Methods like accuracy_score, classification_report, confusion_matrix are used to calculate the performance.

3.3 K-Nearest Neighbor

KNN is a supervised learning algorithm where the result is classified based on the majority vote from its K nearest neighbour category. The algorithm works based on minimum distance from the test data to the training samples to determine the K nearest neighbour. After getting K nearest neighbour a simple majority of them is taken to make prediction of test data. The KNN works as follows: The distance between the test data and all the training samples are calculated. The distance may be calculated by any standard means .Example Euclidean distance. The K nearest neighbour may be included if the distance of the training samples to the query is less than or equal to Kth smallest distance. We then gather a particular feature value of all the nearest neighbours training samples. We take the simple majority of this value as prediction and categorize our new test data.

We use the KNeighborsClassifier in the scikit-learn to implement the KNN. The model requires the number of neighbors as the input parameter. By simply varying the number of neighbors we can find the type of seed. Using this method an accuracy of 90% is achieved.

Class	Precision	Recall	f1-score	Support
1	0.81	0.93	0.87	14
2	1.00	0.86	0.92	7
3	0.95	0.90	0.93	21
Avg / total	0.91	0.90	0.91	42

Table 2: Seed Classification using KNN

3.4 Support Vector Machine

SVM is a linear classifier, uses hyperplane to separate the classes. If our dataset is linearly separable their exists infinite hyperplane to separate the data. For non-linearly separable data's we map the datapoints into high dimensional feature space where they will be linearly separable. As there exists infinite hyperplanes, the one with the largest margin is chosen. The margin is a positive distance from the decision hyperplane. The support vectors are the training patterns that are close to the hyperplane and are the most difficult patterns to classify. Hyperplane is defined by a set of weights and bias. Maximum margin hyperplane for given training set need to be found. 88%

Class	Precision	Recall	f1-score	Support
1	0.69	0.90	0.78	10
2	1.00	0.90	0.95	20
3	0.91	0.83	0.87	12
Avg / total	0.90	0.88	0.89	42

Table 3: Seed Classification using Support Vector Machine

3.5 Decision Tree

The decision tree uses the training data to create a tree and will progressively split the data into smaller subsets in each step. The classification of a particular pattern begins at the root node which checks for a particular property of a pattern. Different branches correspond to different categories. Based on the result to current node, we follow appropriate link to the descendent node. Each leaf node bears a category label and the test pattern is assigned the category of leaf node reached. Building a decision tree using scikit-learn for the wheat seed dataset is achieved with 92% accuracy.

Class	Precision	Recall	f1-score	Support
1	1.00	0.94	0.97	16
2	1.00	1.00	1.00	10
3	0.94	1.00	0.97	16
Avg / total	0.98	0.98	0.98	42

Table 4: Seed Classification using Decision Tree

3.6 Random Forest

Random Forest is a type of ensemble learning method, where a group of weak models combine to form a powerful model. It is a versatile machine learning method capable of performing both classification and regression. In RF, we grow multiple trees instead of single tree. Samples of training data are chosen at random with replacement. This will be used to grow the tree. Each tree is grown to the largest extent possible without any pruning. To classify a new test data, each tree gives a classification and votes for a class. The forest chooses the classification having the most votes. It is a type of unsupervised clustering and can handle large dataset with high dimensions. We got an accuracy of 93% using this method

.

Class	Precision	Recall	f1-score	Support
1	1.00	0.75	0.86	16
2	1.00	1.00	1.00	10
3	0.80	1.00	0.89	16
Avg / total	0.92	0.90	0.90	42

Table 5: Seed Classification using Random Forest

3.7 Naive Bayes

Naive Bayes is based on popular Bayes Decision theory which is a primitive algorithm for Machine learning Techniques. Classification decision here is based on probability. The likelihood, prior probabilities and evidence are used to compute the posterior probability. Evidence is a mere scaling factor for ensuring that the posterior probability sums to one. The category having the highest posterior probability is selected as the resultant category for the particular test data. Naive Bayes classifier assumes that every feature contribute independently to the probability, that is one feature does not affect other. We use GaussianNB classifier of scikit-learn and got an accuracy of 85%.

Class	Precision	Recall	f1-score	Support
1	0.88	0.58	0.70	12
2	0.86	0.92	0.89	13
3	0.85	1.00	0.92	17
Avg / total	0.86	0.86	0.85	42

Table 6: Seed Classification using Naive Bayes

Classifier	**Accuracy**
KNN	90%
SVM	88%
Decision Tree	92%
Random Forest	93%
Naive Bayes	85%

Table 7: Summary of Various Classifiers Accuracy

4. CONCLUSION

This work presents WheatSeed Dataset classification using algorithms such as KNN, Support vector machine, Random forest, Decision Trees and Naives Bayes. Built-in methods for the classifiers are used from the Scikit-learn library. The percentages of accuracy produced by the various algorithms are discussed. The machine learning classifiers can be used for any similar classification problem where we have the features given.

5. REFERENCES

1) M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak," A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek

Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

2) Ahmad Reza Parnian, Reza Javidan , " Autonomous Wheat Seed Type Classifier System " International Journal of Computer Applications (0975 – 8887) Volume 96– No.12, June 2014.

3) Raja Haroon Ajaz , Lal Hussain "Seed Classification using Machine Learning Techniques" Journal of Multidisciplinary Engineering Science and Technology.

4) L.Lin and L.Suhua, "Wheat Cultivar Classifications Based on Tabu Search and Fuzzy C-means Clustering Algorithm", Fourth International Conference on Computational and Information Sciences, pp. 493-496, Aug 2012.

5) M.R. Neuman, E. Shwedy and W. Bushu , "A PC-based colour image processing system for wheat grain grading", International Conference on Image Processing and its Applications, pp. 242-246, Jul 1989.

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

7) 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.

8) 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.