COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS: AN APPLE FRUIT AS CASE STUDY

[1]Chirag Patel, [2]Dr. Vimal K. Chaudhari
[1]Assitant Professor, [2] Assitant Professor
[1] SDHG college of BCA & I.T., Udhna, Surat

Abstract: This paper is aim to classify Apple fruit image on different classifiers: SVM, KNN, Naïve Bayes, Random Forest and Neural Network. There are several methods for fruit classification which are based on color and shape based but fruit has same color and shape which will not enough to identify fruits, proposed method is based on Color, Zone, Area, Centroid, Size, Equvidiameter, Perimeter and Roundness. This features are extracted from the images and provided to classifier models as training and classifying, to train classifiers supervised training approach is used. To compare classifiers Orange canvas and Weka datamining tools are used and found that SVM has highest 95.8% and 87.5% accuracy.

IndexTerms **- Fruit classification, Image classification, Image recognition, SVM, KNN, Random forest, Naïve Bayes, Random Forest, Neural Network**

I. INTRODUCTION

In the past few years there were rapid technology changes arrived which created large amount of image data in various sectors, after introduction of smart phones and increased usage of social media has forces users to create large amount of image data. Detecting and locating of the images is the one of the major factor for user to store image efficiently and to reduce time from system. There are lot of challenges to store image and to retrieve image, to store image it takes lot of storage space on the secondary storage devices so its needs to be compressed for storage and to retrieve image it needs to identify images for that there are various methods are used to recognize images. The traditional image retrieval is based on the keyword annotation. However, there are some difficulties in describing the images by keyword only, because much manual labor is required and the users might give different interpretations depending on their subjectivity. To overcome these limitations, content-based image retrieval (CBIR) has been developed, which exploits the visual contents of the images such as color, texture and shape features [1].

Content based image retrieval is method which will extract visual content of the images automatically from the provided images and recognize images based on fetched visual contents. Extracting visual content from the image is the process of acquiring features from the images, then make analysis of the features is done which required more computation power to process. Feature extraction is the general term for method of constructing combination of the variables to get around problems while still describing the data with sufficient accuracy.

Features which extracted from the images is the subsequence of the measurement of the patterns which transformed to pattern features. Pattern features extracted from the images are assigned to categories or classes, it will identify the images as provided algorithms. Pattern recognition is an important field of computer science concerned with recognizing patterns, particularly visual and sound patterns. It uses methods from statistics, machine learning and other areas. Classification will process the unknown object in query image and will be compared to every sample of the objects that are previously provided to train classifier algorithm.

There are two types of training available for classifiers 1) Supervised training: is the task of inferring a function from labeled training data. The training data consist of a set of training examples. Each example is a pair consisting of an input object (typically a vector) and a desired output value. An algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. 2) Unsupervised training: trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. In proposed system supervised training approach is used for classifiers.

Fruit recognition is the major part of computer vision, there are many attempts are made for fruit recognition among them first was VeggieVision was the first supermarket produce recognition system consisting of an integrated scale and image system with a user-friendly interface [2], Woo Chaw Seng, Seyed Hadi Mirisaee has provided fruit recognition based on colorbased, shape-based and size-based [3]. Shiv Ram Dubey and Anand Singh Jalal has introduces state-of-art colour and texture features and combined them to achieve more efficient and discriminative feature description [4].

In the past various fruit recognition approaches are introduced but most of based on color and shape based of fruits but fruits may have similar color and shapes which will not classify images efficiently so proposed fruit recognition approach is based on Color, Zone, Area, Centroid, Size, Equvidiameter, Perimeter and Roundness features of the fruit images.

This paper provide comparative analysis of the Apple fruit images on SVM, KNN, Random Forest, Naïve Bayes and Neural Network classifiers of machine learning algorithm, to measure performance of the classifiers Orange Canvas and Weka datamining tools are used.

II. METHODOLOGY

The proposed content based image retrieval method for fruit recognition is based on Color, Zone, Area, Centroid, Size, Equvidiameter, Perimeter and Roundness as features of the fruit images. In this study experiments are performed on Apple fruit, 22 images are used for training purpose and for testing 6 images are used.

Matlab is used for image pre-processing and feature extractions for fruit image which will create array of the image features for training images and testing images, which will be exported to Orange canvas and Weka to measure performance of the classifiers [5].

Orange Canvas is the one of the datamining toolbox which will provide facility of the interactive data visualization, visual programming, image analytics, visualizing multiple variables and stacking which will enable user to combine multiple models, in Orange Canvas training data is imported which was exported by Matlab then it passed to different classifier models: SVM, KNN, Random Forest, Naïve bayes and Neural network afterwards test data is imported which was also exported by Matlab and performed testing of model using test and score tool of the orange canvas which will generate performance of model and by using test and score tool, confusion matrix, calibration plot and ROC analysis of the models have been generated.

Weka is collection of machine learning algorithm for datamining. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. It is an open source software issued under GNU general public licenses. Weka is providing Knowledge Flow tool from which comparison of Classification model flow can be created. By using CSVloader tool CSV data of training data imported to weka as exported by Matlab then Class Assigner, Class Value Picker and Cross Validation Fold Maker tool will find classes, select values and validate data which have been provided in the CSVloader tool after parsing of data it passed to the different classifier models: SVM, KNN, Naïve Bayes, Random Forest and Neural Network tool as training data and same process applied to Testing data and passed to the different classifier models. By using classifier performance evaluator tool performance of the model have been calculated and then by using model performance chart will generate Model Performance Chart and by using cost benefit analysis tool cost analysis chart has been generated [6].

III. FEATURE EXTRACTION

To extract features from images, image dataset is taken from the user i.e. path of the Apple images folder. System will read images from the provided path one by one and resize the image to200X200 pixels to make all images same scale then images are converted to RGB to Grayscale to identify the region of the fruits from images, to identify region of the fruit background of the image needs to be subtract, to subtract background thresholding of method is applied because images background are not same for every images.

To perform threshold method to subtract background requires to find threshold of images. Threshold found by identifying edges of the fruit region from the image and by performing dilation operation on the image by using threshold image background will be removed and grayscale image is now converted to binary image and then features are extracted except color feature which extracted before converting image to Grayscale.

- Color:

To identify color of the fruit Mean of the color has been calculated, to find more accurate Mean value of color proposed method is cropping five parts from the images and calculate mean value of the each part and then again find the mean value of all five parts of the images. Parts have been taken from different location of the image i.e. Top, Middle, Bottom, Left and Right.

- Zone:

To identify Zone, binary image portioned into four quadrants then make summation of each quadrant and calculate the percentage of the each quadrant.

- Area:

Area of the fruit will be extracted from the binary image. This will find the fruit portion of the binary image. Area will be found using analyzing number of elements of rows pixel of the binary image and provided as feature to classifier.

- Centroid:

Centroid of the fruit area is the average Mean pixels of rows and columns of the fruit area of binary image. Average mean of the rows is the Major axis length value of the centroid and average Mean pixels of columns is Minor axis length value of the centroid and provided as feature to classifier.

- Size:

Size of the fruit is height and width of the fruit area which have been calculated after converting image into binary image.

- EquviDiameter:

EquviDiameter is to find the diameter of the fruit image region with same area as region which will be found in Binary image. To calculate EquviDiameter sqrt(4 *Area / pi) formulation is used.

- Perimeter:

Perimeter is the distance of the region of the fruit. It computes the perimeter by calculating the distance between each adjoining pair of pixels around the border of the region, it will extracted from the fruit region and provided to the classifier.

- Roundness:

Roundness is dominated by the shape's gross features rather than the definition of its edges and corners, or the surface roughness of a manufactured object. It will be calculated using (4 * Obj_area * pi) / Per .^ 2 formulation.

IV. CLASSIFIERS

After extracting features from the exported data from both Orange Canvas and Weka tools data is provided to different models: SVM, KNN, Naïve Bayes, Random forest and Neural Network classifiers to train classifier model and then measure performance of the model on provided testing data.

- Support Vector Machine

Support vector machines (SVMs) were originally designed for binary classification. As it is computationally more expensive to solve multiclass problems, comparisons of these methods using large-scale problems have not been seriously conducted. Especially for methods solving multiclass SVM in one step, a much larger optimization problem is required so up to now experiments are limited to small data sets. Decomposition implementations for two such "all-together" methods. We then compare their performance with three methods based on binary classifications: "one-against-all," "one-against-one," and directed acyclic graph SVM (DAGSVM). Our experiments indicate that the "one-against-one" and DAG methods are more suitable for practical use than the other methods [7].

- K – Nearest Neighborhood

Basically the algorithm works by comparing a given test tuple with training tuples that are similar to it. K-nearest neighborhood classifier searches the k training samples that are closest to unknown sample. Closeness is defined in terms of Euclidean distance which can be computed by equation. In K-nearest neighborhood algorithm, classification results mainly depend on value of k which is a design parameter and generally is obtained empirically [8].

- Naïve Bayes

The naïve Bayesian classifier makes the assumption of class conditional independence, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. This simplifies computation. When the assumption holds true, then the naïve Bayesian classifier is the most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables [9]

- Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [10].

- Neural Network

A neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units. It is difficult for humans to interpret the symbolic meaning behind the learned weights and of "hidden units" in the network [11].

V. RESULTS

To identify performance of the classifiers 22 apple images are used for training purpose and 6 images are used for testing.

Sampling type: No sampling, test on testing data
Target class: Apple

Scores

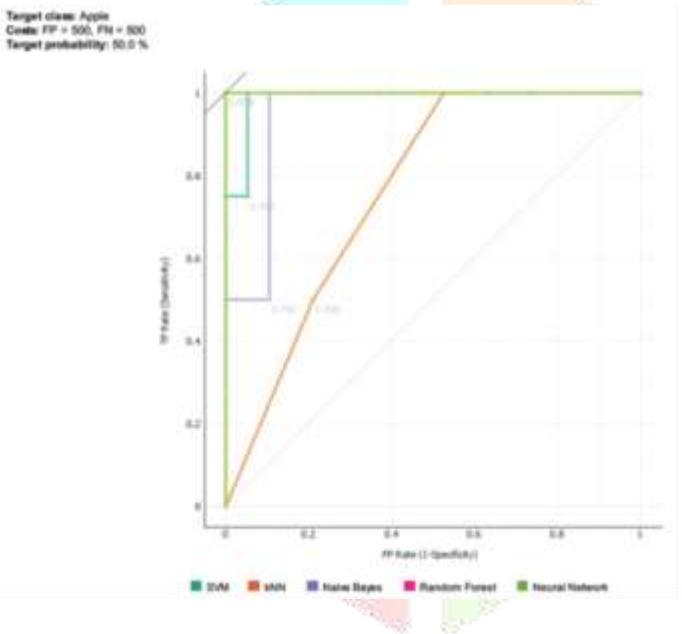
Method	AUC	CA	F1	Precision	Recall
SVM	0.926	0.958	0.909	0.833	1.000
kNN	0.600	0.625	0.308	0.250	0.400
Naive Bayes	0.874	0.792	0.545	0.500	0.600
Random Forest	0.700	0.875	0.571	1.000	0.400
Neural Network	0.716	0.750	0.000	0.000	0.000

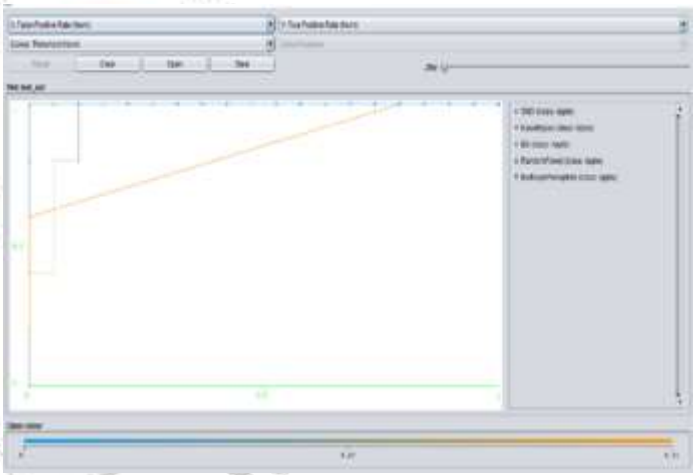
Fig.1 Detail Accuracy of Classifiers in the Orange Canvas

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
SVM	0.875	0.025	0.900	0.875	0.865	0.855	0.946	0.805
KNN	0.833	0.031	0.856	0.833	0.835	0.807	0.906	0.754
Naïve Bayes	0.583	0.078	0.561	0.583	0.547	0.498	0.964	0.895
Random Forest	0.750	0.036	0.806	0.750	0.751	0.724	0.976	0.876
Neural Network	0.750	0.039	0.765	0.750	0.745	0.714	0.946	0.846

Fig.2 Detail Accuracy of Classifiers in the Weka

Above in the Fig.1 and Fig.2 shows detailed accuracy report of the classifiers in the Orange Canvas and Weka tool box. Detailed accuracy shows that in Orange canvas SVM has 95.8%, KNN has 62.5%, Naïve Bayes has 79.2%, Random forest has 87.5% and Neural Network has 75% of accuracy and in Weka SVM has 87.5%, KNN has 83.3%, Naïve Bayes has 58.3%, Random Forest has 75% and Neural Network has 75% of accuracy.





Above Fig.3 and Fig.4 shows ROC analysis diagrams of Apple fruit in Orange canvas and Weka. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning.

VI. CONCLUSION

In this research Apple fruit categorization is done on different classifiers: SVM, KNN, Random Forest, Naïve bayes and Neural network and measured performance of the classifiers based on Color, Zone, Area, Centroid, Size, Equvidiameter, Perimeter and Roundness. Both Orange Canvas and Weka shows that SVM has highest accuracy of 95.8% and 87.5% of accuracy and ROC analysis also shows SVM provides an efficient performance over other classifiers.

REFERENCES

[1] Jung Won Kwaka, Nam Ik Chob, Relevance feedback in content-based image retrieval system by selective region growing in the feature space, Signal Processing: Image Communication 18 (2003) 787–799

[2] R. M. Bolle J. H. Connell N. Haas R. Mohan 6. Taubin "Veggievision: A produce recognition system."Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on. IEEE, 1996.

[3] Woo Chaw Seng and Seyed Hadi Mirisaee "A New Method for Fruits Recognition System" 2009 International Conference on Electrical Engineering and Informatics

[4] Shiv Ram Dubey and Anand Singh Jalal "Fruit and vegetable recognition by fusing colour and texture features of the image using machine learning" International Journal of Applied Pattern Recognition Vol. 2, No. 2, 2015

[5] Ralph Gonzalez (Author),Richard Woods (Author), Steven Eddins, Digital Image Processing Using MATLAB

[6] Ian H. Witten, Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques

[7] Chih-Wei Hsu and Chih-Jen Lin, A Comparison of Methods for Multiclass Support Vector Machines, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 2, MARCH 2002

[8] Deepika Shukla and Apurva Desai, Recognition of Fruits Using Hybrid Features and Machine Learning, 2016 International Conference on Computing, Analytics and Security Trends

[9] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 315

[10] LEO BREIMAN, Random Forests, Machine Learning, 45, 5–32, 2001

[11] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 327