International Journal of Scientific Research in Multidisciplinary Studies
Vol.5, Issue.12, pp.12-14, December (2019)

Research Paper
E-ISSN: 2454-9312
P-ISSN: 2454-6143

Efficiency of Feature Selection on Banana Classification

A. Anushya

Dept. of Computer Applications, Manonmaniam Sundaranar Univesity, India

Author: anushya.alpho@gmail.com

Available online at: www.isroset.org

Received: 10/Dec/2019, Accepted: 15/Dec/2019, Online: 31/Dec/2019

***Abstract*—**This research compares the impact of feature selection in classification accuracy of Decision Tree on French Plantain (Nendran) dataset. At first, a five class home-made database is created and classification algorithm, Decision Tree is employed on dataset before and after feature selection via Rough set. The experiments are carried out on MATLAB. Results reveals that classification after feature selection produces 69% of classification accuracy which is enhanced than the before.

Keywords—Classification, Decision Tree, C4.5, Feature Selection, Relative Reduct

I. INTRODUCTION

French Plantain (Nendran, Ethakkai, Myndoli and Rajeli) is a popular plantain variety of Kerala and Tamil Nadu. The fruits are large and plumpy with long pedicel. Fruits turn buff yellow on ripening [1]. To classify the Banana is tough task. There are number of researches are conducted by researches. Some of them is described as follows. Ridhuna Rajan et.al., [2] proposed an approach to identify the infected parts of fruit using Improved K-means clustering and segmentation. It detect the fruit parts faster and it is less time consuming. Emny Harna Yossy et.al., constructed a system using C language , compute vision and Artificial Neural Networks to detect the mango that has been ripe or not using color attribute besides 94% of accuracy is attained [3]. Nandhan Thor et.al., conducted the experiment with Banana images and extracted color features and then clustered for determining the banana ripeness states. Decision tree is applied to classify fruit shelf-life and achieved 52% of accuracy [4]. Jonatha et.al., [5] applied Image processing methods by image filtering, segmentation and feature extraction for citrus fruits. Support Vector Machine used to classify the quality of citrus fruits and accuracy was gained 97.3%. Artificial Neural Networks based frame work using color, brown spots development and Tamura features to classify the banana images by Fatima.M.A.Mazen et.al.,. To verify the result, comparative analysis was done by classification algorithms such as Support Vector Machine, Naïve Bayes, K-Nearest Neighbor and Decision Tree. Overall 100% class recognition accuracy is obtained [6].

From the above literature, the research works were dome with images and supported in Image processing techniques. This paper deals Data mining techniques to classify the French Plantain. Data Mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, forecasting and estimation.

The rest of the paper is organized as follows. Data source is described in section II. Research methodology is depicted in Section III, and Section IV obviously displays the result analysis from the experiments and Section V concludes the paper while including argument on finding of the research and on probable future work.

II. DATA SOURCE

The database comprised of 210 French Plantain instances with different ripening stages such as unripe (green), mid ripe (yellow with green), ripe (golden yellow), over ripe (yellow with black) and defected. The dataset consists of 5 classes as 50 unripe, 50 mid ripe, 50 ripe, 50 over ripe and 10 defected bananas. For this study the banana were collected from Kaniyakuamri District, Tamil Nadu, India and prepare the dataset with 7 conditional attributes such as length, width, Peel of skin color, pull tab color, pull tab width, surface and whether splitted or cut and one decision attribute. Sample dataset with 10 instances is depicted in Table 1.

Table 1. Sample dataset with 10 instances

Length	Width	Skin color	Pull tab color	Pull tab width	Surface	Split	Class
22	5	Yellow with green	Green	1.5	Smooth	No	Mid ripe
21	5.5	Yellow with green	Green	1.0	Smooth	No	Mid ripe
17	3.5	Golden yellow	Green	0.75	Smooth	No	Ripe
21.5	4	Yellow with black	Black	1.0	Smooth	No	Over ripe
19	5.5	Yellow with black	Green	2	Smooth	No	Ripe
18	5.5	Golden yellow	Green	2	Smooth	No	Ripe
16	5	Yellow with black	Black	1	Rough	No	Over ripe
17	4	Green	Green	1.5	Smooth	No	Un ripe
16.5	5	Golden yellow	Black	0.5	Smooth	Yes	Defected
17.5	4.5	Yellow with black	Black	0.75	Smooth	No	Over ripe

III. RESEARCH METHODOLOGY

Initially Decision Tree is used for classification of real French Plantain dataset. Then feature selection via Relative Quick Reduct is applied and reduced features are acquired. Again classification employed on reduced features. Comparison of with and without feature selection is analyzed. Techniques used in this works such as classification and feature selection are explained in detail.
Classification

Rule discovery is an important data mining task since it generates a set of symbolic rules that describe each class or category in a natural way. Classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, as determined by one of the attributes, called the class label attribute. In the context of classification, data tuples are also referred to as samples, examples, or objects. In this work, Decision Tree is used as a classification algorithm. Next, Steps involved in Decision Tree is explained.

Decision Trees

Decision trees [7, 8] are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Decision tree is a classifier in the form of a tree structure and has decision node and leaf node. Decision node specifies a test on a single attribute, Leaf node indicates the value of the target attribute, Arc or edge is split of one attribute and Path is a disjunction of test to make the final decision. Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. Uses a tree structure to model the training set and classifies a new record following the path in the tree and inner nodes represent attributes and leaves nodes represent the class. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution. The top most node in a tree is the root rode. The most popular Decision tree algorithms are CART, CHAID and C4.5. This section outlines C4.5 algorithm, by first introducing the basic methods of its predecessor, ID3 algorithm. The basic algorithm of C4.5 is discussed below.

C 4.5 Algorithms

C4.5 builds decision trees from a set of training data using the concept of information entropy. The decision tree prescription for synthesizing an efficient decision tree can be stated as follows:

Step 1: Calculate initial value of entropy.

Step 2: Select that feature which results in the maximum decrease in entropy (gain in information), to serve as the root node of the decision tree.

Step 3: Build the next level of the decision tree providing the greatest decrease in entropy.

Step 4: Repeat Steps 1 through 3. Continue the procedure until all subpopulations are of a single class and the system entropy is zero.

Feature Selection

Volumetric features can increase system density and may not for all time lead to higher prediction accuracy. Though, these features are not independent and may be correlated. A bad feature may seriously disgrace the performance. Thus, selecting a subset of good features is important. Features are selected by a learning algorithm during the training phase. The selected features are used as a model to describe the training data. Consequently smaller quantity of features can reduce the computational outlay, which is vital for real-time applications. Also it may lead to better classification accuracy due to the finite sample size effect. In this work, we use Relative Reduct Algorithm to select as few features as possible to describe the training data effectively. The steps of Relative Reduct Algorithm [9] is shown below.

Relative Reduct (C, D)

1) R<- C
2) ∀ a ∈c
3) if (K R - {a} (D) ==1)
4) R ← R-{a}
5) Return R

The RelativeReduct reduces the feature for dataset with proliferation of features. Then classification of dataset with reduced features will be conducted. The experimental results and its comparison are specified in next section.

IV. RESULT ANALYSIS

First Decision Tree has been used in MATLAB on the real French Plantain data. Totally 210 instances with 7 conditional attributes (length, width, Peel of skin color, pull tab color, pull tab width, surface and whether splitted or not) and 1 decision attribute with 5 values such as unripe, mid ripe, ripe, over ripe, splitted. Then Feature selection via Relative Reduct is applied and the reduced features are obtained. The reduced features are length, width, Peel of skin color, surface and whether splitted or not. Again classification by Decision Tree with reduced features by Relative Reduct. The results are visibly presented in Table 2.

Table 2. Effectiveness of Feature Selection

Feature selection	Dataset	Instances	Number of instances predicted in				
			Unripe	Mid ripe	Ripe	Over ripe	Defected
Before	Training	147	32	35	39	28	5
	Testing	63	18	15	11	22	5
After	Training	147	35	33	39	31	7
	Testing	63	15	17	11	19	3

There are two different classes acting as the data source. The dataset is divided into the ratio of 70:30, where 70% is for training and 30% is for testing. We used 147 instances in training phase among these 35 are unripe, 35 are mid ripe, 35 are ripe, 35 are overripe and 7 defected. Also 63 records are used in testing phase. 15 are unripe, 15 are mid ripe, 15 are ripe, 15 are overripe and 3 defected in testing phase. From the above table, it is absolutely illuminated the outcomes of Feature selection. The classification accuracy obtained 66% and 69% before and after feature selection respectively.

V. CONCLUSION AND FUTURE SCOPE

In this paper, a sub problem of Data mining, classification via Decision Tree is discussed and importance of feature selection is encountered. This paper manifestly expressed the efficacy of feature selection on French Plantain dataset. The concert of decision tree with out and with feature selection are compared and accuracy obtained by Decision Tree after feature selection is 69% which is higher than before. In near future the research will be conducted by images of bananas with same techniques and analyze effectiveness is compared by data vs. images.

REFERENCES

[1] N.S. Lele , "*Image Classification Using Convolutional Neural Network*", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.3, pp.22-26, 2018.

[2] Ridhuna Rajan et.al., "Analysis and Dectection of Infected Fruit part using Improved K-means clustering and segmentation Techniques", IOSR Journal of Computer Engineering (IOSR), Pp:**37-41, 2015**.

[3] Emny Harna Yossy et.al., "Mango fruit sortation system using Neural Networks and computer vision", Procedia Computer Science, Volume: **116**, Pp:**599-603, 2017**.

[4] Nandhan Thor et.al., " Applying Machine Learning clustering and classification to predict Banana ripeness states and shelf life", International Journal of Advanced food science and technology, Cloud Publications, Volume **2, issue :1,**pp**:20-25, 2017**.

[5] Jonatha et.al., " Citrus Friut quality classification using support vector machine", International Journal of Advanced Engineering Research and Science, Vol:**6,** Issue**:7, 2019**.

[6] Fatima.M.A.Mazen et.al., "Ripeness classification of bananas using Artificial Neural Networks", Arabian Journal for science and Engineering, **2019**.

[7] Raj Kumar et.al, "Classifiction algorithms for Data Mining", A Survey,International Journl of Innovations in Engineering and Technology,Vol.**1,** Issue **2 ,2012**.

[8] Veronical, S.Moetini, "Towards the use of C4.5 Algorithm for classifying Banking Data set", Integral, Vol.**8.** No**.2, 2003**.

[9] K.Thangavel et.al., " Dimensionality reduction based on rough set theory: A review", Applied softcomputing, volume **9, issue 1, 2009.**

AUTHORS PROFILE

Dr.A.Anushya pursed Bachelor of Science from Mother Teresa University, India in 2006. Master of Computer Applications from Bharathidhasan University in 2009 and Ph.D. Computer Applications from Manonmaniam Sundaranar Univesity, India inn2016.. She is previously working as Assistant Professor in Coolege of Computer Science and Software Engineering, University of Hail, Hail, Kingdom of Saudi Arabia. She has published more than 15 research papers in reputed international journals including IEEE and also available online. Her main research work focuses on Data Mining, image mining and soft Computing. She has 5 years of teaching experience and 10 years of research experience.

