Classification of Jujube Fruits Using Different Data Mining Methods

Ali Rahimi, Ahmad Banakar*, Hemad Zareiforoush, Mehrdad Beygvand, Mehdi Montazeri

Deptartment of Agricultural Machinery Engineering, Faculty of Agriculture, Tarbiat Modares University, P.O. Box 14115-111, Tehran 14114, Iran
Email: ah_banakar@modares.ac.ir

Abstract: In the current study, the qualitative grade of jujube fruit was evaluated using machine vision and classifying techniques. Images from four different classes of jujubes (G1, G2, G3, and G4), representing the quality grades of jujube fruits, were acquired using a color CCD camera. After pre-processing and segmentation of images, 57 features including five from size and shape, four from texture, and 48 from color information were extracted. To select the best features for grading of the jujubes, correlation-based feature selection was used. It was revealed that 13 features surpassed the other features in quality classification. Afterwards, four different data mining-based techniques including artificial neural networks, support vector machines, decision trees and Bayesian Networks were used to classify jujubes. Results of validation stage showed that artificial neural network with 13-7-4 topology had the highest classification accuracy, 98.61%. After artificial neural network, support vector machine with polynomial kernel function (95.91%), Bayesian Network with Hill Climber search algorithm (95.22%), and decision tree with J48 algorithm (95.14%) had higher accuracy, respectively. Results of this research can be adapted for developing an efficient system for fully automated sorting of jujube fruits.
[Rahimi A, Banakar A, Zareiforush H, Beygvand M, Montazeri, M. **Classification of Jujube Fruits Using Different Data Mining Methods.** *Researcher* 2014;6(5):52-61]. (ISSN: 1553-9865). http://www.sciencepub.net/researcher. 9

Key words: Image processing; Data mining; Jujube

1. Introduction

Jujube or Chinese date (Zao or Hongzao in Chinese), *Ziziphus jujube* Mill. (Z *sativa* Gaetn., Z. *vulgaris* Lam), belongs to the family of Rhamnaceae and order Rhamnales. It is a native fruit and medicinal plant of China with a very distinct characteristic of producing deciduous bearing branches. Jujube is now commercially produced in China, South Korea and Iran but grown mainly for ornamental or research purposes in many other counties. It is grown in the temperate and subtropical areas of the Northern Hemisphere, especially the drier parts of Southern Khorasan province, Iran. The fruits are mainly consumed fresh, dehydrated, or processed into candy, jam, juice, rich in nutrition, easy to manage and has multiple uses and fits for long-term intercropping systems.

Currently, jujube grading is performed manually based on the product important quality indices. Nevertheless, manual grading is costly and unreliable, as human inspectors' decisions are often incompatible with each other. Another way for jujube sorting is mechanical method which is carried out only based on shape properties of the product. It is evident that use of such methods cannot guarantee the precise sorting of jujube because the other important quality indices such as decay and water loss cannot be controlled. In this regard, utilization of new technologies such as machine vision and artificial intelligence can be a suitable solution for automated inspection of jujube. It has been stated that utilization of the machine vision and artificial intelligence can result in increased quality of the product, abolish inconsistent manual evaluation, and reduce dependence on available manpower (Li *et al.*, 2009). Nowadays, Computer vision systems are being used increasingly in the food industry for quality assurance purposes. Several studies have been performed on development of machine vision-based techniques for automatic classification of food products. (Qi *et al.*, 2011) presented an on-line machine vision system for shape and size detection of Hami big jujube. To identify the quality defects, they used many methods such as application of wavelet denoising, threshold segmentation, and morphological processing methods. A Hierarchical Clustering Analysis method was applied for validation of the recognition and to determine the classification accuracy. They reported that recognition rate could reach 83%. (Jiang and Zhou, 2013) used machine vision technology for size detection of dried jujube. They reported that the size detection accuracy using the developed system could reach 80-85%. Computer vision has been used for such tasks as shape and variety classification, quality grading and defect detection. Defect segmentation on Golden Delicious apples was carried out by CMV (Leemans *et al.*, 1998). A colour model developed was used as a standard for comparison with sample images. The developed algorithm gave satisfactory results with well-contrasted defects, however two

further enhancements following segmentation were required to improve accuracy. Zayas et al. (1996) found that the physical characteristics of wheat could be used as the basis for the development of an objective wheat classification method. Using computer vision and crush force features, differentiation rate between hard and soft wheat was 94% for the varieties tested. Yu et al., (2012) reported the usefulness of the least square support vector machine (SVM) for raisin classification. They used color and texture features (obtained from the raisin images) to classify the golden seedless raisins into four classes based on color, shape and degree of wrinkles. Results indicated the highest classification rate by SVM was about 95%. Application of classification methods based on the data mining techniques is an efficacious tool for realizing accurate classifier models (Hu *et al.*, 1998, Al Ohali, 2011, Khadem, 2013). The data mining encompasses decision trees, artificial neural network (ANN), genetic algorithm (GA), fuzzy sets, expert systems, etc. Data mining includes several theories and approaches which, despite being different from one another, have two common denominators: (i) the non symbolic representation of pieces of knowledge and (ii) ‘‘bottom up’’ architecture where the structures and paradigms appear froman unordered beginning. Different powerful algorithms exist for proper selection and extraction of defining features as well as training various data mining models to adapt difficult input–output mappings (Kirkos *et al.*, 2008, Omid *et al.*, 2009, Omid *et al.*, 2010). Existence of undesired factors during the packaging process results in significantly reduction in the final price of the package. These factors can be color variation and existence of the foreign materials such as unseparated tails, leaves and etc. These are visual factors which consumers consider them when buying the product. Therefore, it is necessary to eliminate undesired fruits before packaging by use of automated sorting systems. The objective of this research was to develop a novel computer vision-based algorithm along with an appropriate data mining-based technique to classify jujube fruit based on the product visual features.

2. Materials and methods

The proposed machine vision system for grading of jujube fruits is shown in Fig. 1. In the developed system, first, images of jujubes are captured. The image processing operations are then executed to eliminate unwanted noises from images. After jujube segmentation, the primary feature vector is created according to some shape, size and color features. In order to have a good classification, it is necessary to prepare a good input vector. Therefore, the primary extracted features are subjected to a correlation-based feature selection procedure to select the better features. Finally, the best classifier is selected for jujubes grading by examining four commonly used data mining methods. The whole applied methodology is described in the following sections.

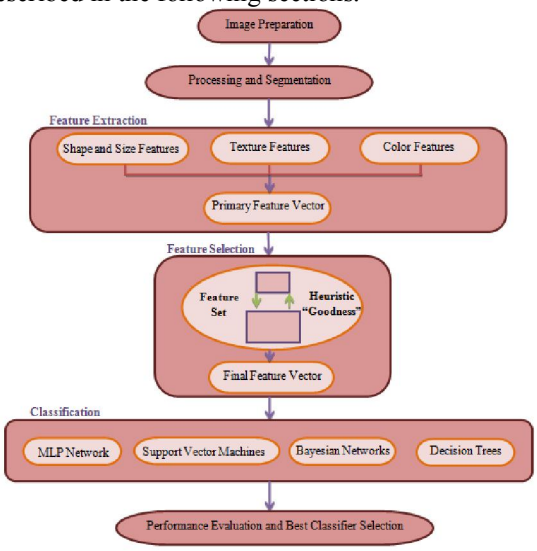

Fig. 1- Proposed methodology for jujube sorting (Mollazade *et al.*, 2012).

2.1. Image acquisition

A machine vision system was developed to acquire images of jujube fruit (Fig. 2). The proposed system consisted of a color CCD camera (Hi-Peak Model 565S, China) equipped with a CS lens mount (3.5–8 mm focal length, 480 vertical TV lines resolution), a video capture card (Pinnacle 510-USB with a resolution 720H × 576V), a personal computer (PC) for image display and acquisition, and an appropriate illumination unit. The CCD camera was placed about 15 cm above the samples and powered by a 12 VDC power supply.

Fig. 2- Image acquisition setup.

In order to provide uniform illumination, strip LED lights were used above the samples. A white cardboard was used as a background surface to simplify the segmentation process. In order to have a uniform illumination condition and to eliminate the environmental noises, the imaging chamber was covered by a black cover. During image acquisition, signals from samples were captured by the camera, digitized and transferred to the PC using the capture card, and stored on the PC in RGB color space.

To capture, record and process the acquired images, a script was written in MATLAB R2010a version (MathWorks, 2010). Before image acquisition, jujubes were manually separated from each other and then were classified by experts into four classes according to the standard provided by Institute of Standards and Industrial Research of Iran (ISIRI, 1990). The classes included grade one jujube (G1), grade two jujube (G2), grade three jujube (G3), and grade four jujube (G4). According to the size of the fruits, a certain number of jujubes were manually placed under camera at each test so that there was no contact between the fruits. For each class, images of 50 fruits were captured. Totally, images of 200 jujubes were obtained from all classes. A sample of the captured images is presented in Fig. 3.

Fig. 3- Sample of captured images: A) Grade 1 jujubes labeled as G1; B) Grade 2 jujubes labeled as G2; C) Grade 3 jujubes labeled as G3; D) Grade 4 jujubes labeled as G4 (Images are shown in the same scale)

2.2. Processing and segmentation

This stage includes the operations used to prepare the images before feature extraction. This is an important process because the result of classification significantly depends on the success of system designer in implementing appropriate process on the images. In this study, the image processing stage consisted of eliminating shadows of fruit, removing background noise, and separating each fruits from the others in the image. In order to separate jujubes from the background, a global threshold was applied on the images using Otsu's method (Gonzalez *et al.*, 2004). Otsu is a histogram based thresholding method in which the normalized histogram is considered as a discrete probability density function (PDF). Otsu's method selects the threshold value k that maximizes value of G based on the following equation (Mollazade *et al.*, 2012):

$$G = P_j(I_j - I_T)^2 + P_b(I_b - I_T)^2 \quad (1)$$

Where P_j is the proportion of pixels of jujubes, P_b is the proportion of pixels of background, I_j is the mean gray value of jujubes, I_b is the mean gray value of background, and I_T is the mean gray value of whole image. The threshold value is converted to a normalized value between 0 and 1. In this study, the threshold value was obtained as 0.375. After thresholding, the segmented images were converted into binary images and then, in order to eliminate shadows surrounding the fruits, the images were subjected to an erosion operation. By performing the trial and error procedure, it was revealed that the effect of shadows can be entirely eliminated by once application of a circle structure with the radius of four pixels. Finally, to carry out feature extraction, images of jujubes were labeled using the developed MATLAB script to remain only one jujube with a specific label in each segmented image (Fig. 4).

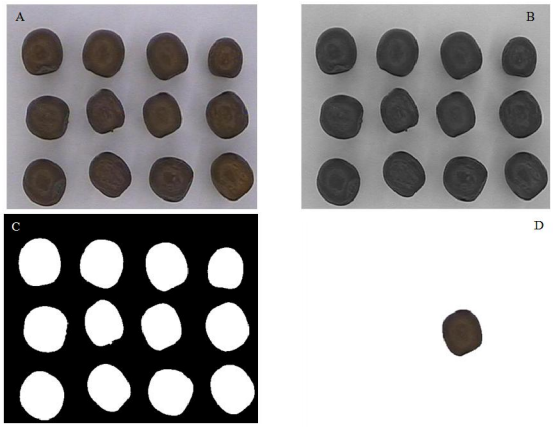

Fig. 4- Sample of jujubes image preprocessing and segmentation operations: A) Original image in RGB color space; B) B channel of original RGB image; C) Image in binary mode after thresholding method; D) A segmented jujube after preprocessing operations.

2.3. Feature extraction

There are many defining features in image processing problems to describe the objects. The feature analyses of jujubes included extraction of color, shape and size features. Totally, 57 features, including five from size and shape, four for texture, and 48 from color information were extracted for each jujube. Table 1 shows the complete list of computed features.

Table 1- Shape, texture and color features of jujube fruits measured by image analysis

	Shape and size features					
Feature	Major Axis Length	Minor Axis Length			Equivalent Diameter	Perimeter	Solidity
Feature No.	F-1	F-2			F-3	F-4	F-5
	Texture features					
Feature	Contrast	Energy			Correlation	Homogeneity
Formula	$\sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{ij}$	$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}^{\ 2}$			$\sum_{i=1}^{k}\sum_{j=1}^{k}(i-j)^2 p_{ij}$	$\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{p_{ij}}{1+\lvert i-j\rvert}$
Feature No.	F-6	F-7			F-8	F-9
	Color features in RGB Space*					
Feature	Mean [μ= $\frac{\sum_{x=0}^{z-1} h(x)}{z}$] for:					
	R	G	B	$R/(R+G+B)$	G/(R+G+B)	B/(R+G+B)
Feature No.	F-10	F-11	F-12	F-13	F-14	F-15
Feature	Mean [μ= $\frac{\sum_{x=0}^{z-1} h(x)}{z}$] for:					
	R-G		G-B		R-B	
Feature No.	F-16		F-17		F-18	
Feature	Variance [σ= $\frac{\sum_{x=0}^{z-1}(h(x)-\mu)^2}{z}$] for:					
	R	G	B	$R/(R+G+B)$	G/(R+G+B)	B/(R+G+B)
Feature No.	F-19	F-20	F-21	F-22	F-23	F-24
Feature	Variance [σ= $\frac{\sum_{x=0}^{z-1}(h(x)-\mu)^2}{z}$] for:					
	R-G		G-B		R-B	
Feature No.	F-25		F-26		F-27	
Feature	Skewness [s= $\frac{\sum_{x=0}^{z-1}(h(x)-\mu)^3}{z\sigma^3}$] for:					
	R	G	B	$R/(R+G+B)$	G/(R+G+B)	B/(R+G+B)
Feature No.	F-28	F-29	F-30	F-31	F-32	F-33
Feature	Skewness [s= $\frac{\sum_{x=0}^{z-1}(h(x)-\mu)^3}{z\sigma^3}$] for:					
	R-G		G-B		R-B	
Feature No.	F-34		F-35		F-36	
Feature	Kurtosis [k= $\frac{\sum_{x=0}^{z-1}(h(x)-\mu)^4}{z\sigma^4} - 3$] for:					
	R	G	B	$R/(R+G+B)$	G/(R+G+B)	B/(R+G+B)
Feature No.	F-37	F-38	F-39	F-40	F-41	F-42
Feature	Kurtosis [k= $\frac{\sum_{x=0}^{z-1}(h(x)-\mu)^4}{z\sigma^4} - 3$] for:					
	R-G		G-B		R-B	
Feature No.	F-43		F-44		F-45	
	Color features in HSV Space					
Feature	Mean [μ= $\frac{\sum_{x=0}^{z-1} h(x)}{z}$] for:					
	H		S		V	

Feature No.	F-46	F-47	F-48
	Color features in L*a*b* Space		
Feature	Mean [μ= $\frac{\sum_{x=1}^{z} h(x)}{z}$] for: L*	a*	b*
Feature No.	F-49	F-50	F-51
	Color features in YCbCr Color map		
Feature	Mean [μ= $\frac{\sum_{x=1}^{z} h(x)}{z}$] for:		
	Y	Cb	Cr
Feature No.	F-52	F-53	F-54
	Color features in NTSC System		
Feature	Mean [μ= $\frac{\sum_{x=1}^{z} h(x)}{z}$] for:		
	Yi	I	Q
Feature No.	F-55	F-56	F-57

* h(x) is the grey level of pixels in the image with a pixel position of x , x can take any value between 1 and $z= m \times n$, where m and n are number of rows and columns of the image matrix, respectively.
* μ, σ, S, and K are the *Mean*, *Variance*, *Skewness*, and *Kurtosis* of image pixels, respectively.

2.4. Feature selection

The next step after feature extraction is to select the superior features from the extracted feature vector. In data mining problems, selection of appropriate feature vectors is important because these are the only data fed to the classifiers. Selection of the best features is one of the key factors in improving each classifier performance. In this regard, it is essential to have a sufficient feature vector. Nevertheless, features that do not improve classification accuracy should be removed from the feature vector. Several techniques are available for feature selection. The most commonly used applicable approaches for classification purposes include Principal Component Analysis (PCA), Correlation-based Feature Selection (CFS), factor analysis, and sensitivity analysis (Omid *et al.*, 2010). CFS is one of the prominent data mining methods to rank the relevance of features. It uses a search algorithm along with a function, Pearson's correlation equation, to evaluate the merit of feature subsets (Mollazade *et al.*, 2012). The heuristics by which CFS measures the goodness of feature subsets takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them (Hall, 1999). In the current study, "Best First" procedure was chosen as the search algorithm. Best first algorithm searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. The level of backtracking done can be controlled by setting the number of consecutive non-improving nodes allowed. Best first may start with an empty set of attributes and search forward, or it may start with a full set of attributes and search backward, or it may even start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point) (Witten and Frank, 2005). The mentioned algorithm was implemented on the extracted features of jujubes using CfsSubsetEval attribute evaluator in WEKA software (Hall *et al.*, 2009). The CfsSubsetEval algorithm evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. After feature selection operation, the size of feature vector showed a reduction from 57 features to 13 features, including 4 for size and shape, 3 for texture, and 6 for color features. The selected features were: F-1, F-3, F-4, F-5, F-6, F-7, F-8, F-12, F-21, F-35, F-36, F-46, and F-53 (Table 1).

2.5. Intelligent classification of jujubes

Classification was the last stage of the jujube grading process. Generally, classification is the process of training to assign a sample to pre-determined classes. The aim of classification was to find a rule based on the selected features or training elements, which allowed assigning each jujube to any of probable classes. Since the classification process contains training, cross-validation, and testing stages, the data set had to be divided into three parts: training set, cross-validation set, and testing set. The training set was used to train the classifier; whist cross-validation set was utilized to prevent the

overtraining and the testing set was employed to test the validity of the classifier. In this study, 60% of data set (120 samples) was randomly selected as training set, 20% (40 samples) for cross-validation, and the remaining 20% of data set (40 samples) was used as testing set. Several strategies can be implemented for the classification process. Most of them are categorized as data mining-based techniques. Here, to find the best classifier for jujube grading, four different data mining techniques were evaluated using WEKA software (Hall *et al.*, 2009). Each of the utilized techniques is described in the following sections.

2.5.1. Artificial neural network

In computer science and related fields, artificial neural networks are machine learning models inspired by animals' central nervous systems that are capable of simulating the behaviour of human brain. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network (Karray and De Silva, 2004). One of the most common types of artificial neural network for classification purposes is Multilayer perceptron (MLP) (Omid *et al.*, 2010, Mollazade *et al.*, 2012). In general, MLPs consist of three main layers: input layers, hidden layers, and output layer. The layers belong to the class of feedforward networks, meaning that the information passes through the network nodes only in the forward direction. In order to classify the jujubes, the MLP model was trained using backpropagation algorithm. This algorithm calculates the weights of the activation function for each neuron (Karray and De Silva, 2004). In the feedforward networks, error minimization can be performed using a number of procedures including gradient descent, gradient descent with a momentum (Omid *et al.*, 2009), Levenberg–Marquardt (Omid *et al.*, 2010), conjugate gradient, and etc. In this research, the gradient descent with a momentum approach was used for error minimization with the momentum coefficient of 0.2 (Mollazade *et al.*, 2012).

2.5.2. Support vector machine

In machine learning, support vector machines are supervised learning systems based on the statistical learning theory that explore data and recognize patterns in classification and regression analysis problems. An support vector machine model is a representation of the samples as points in space, mapped so that the samples of the distinct classes are separated by a clear boundary which is as wide as possible. In this approach, the optimal boundary, known as hyperplane, of two sets in a vector space is obtained independently on the probabilistic distribution of training vectors in the set. The hyperplane locates the boundary that is as far as possible from the nearest vectors to the boundary in both sets. The vectors situated near the hyperplane are called supporting vectors (Ndehedehe, 2014). If the space is not linearly separable, there may be no separating hyperplane to distinguish. In such cases, a kernel function may be used to solve the problem. The kernel function evaluates the relationships within the data and makes complex divisions in the space (Vapnik, 2000).

2.5.3. Decision tree

Decision trees are extremely useful data mining tools. These are a type of machine learning classifiers in which a divide-and-conquer approach results in a style of representation called tree (Mollazade *et al.*, 2009, Omid, 2011). Decision trees are organized so that at each layer of the tree one class is rejected. The last remaining class at the bottom of the tree is considered as the designated class. The outgoing branches of each node correspond to possible outcome of the test at that node. There are a large number of decision tree algorithms introduced completely in the machine learning and applied statistic literatures. In the current research three different decision tree induction algorithms were used for classification of jujubes. The algorithms were namely J48 (C4.5 decision tree learner) algorithm (Mollazade *et al.*, 2009, Omid, 2011), REP (reduced-error pruning), and LMT (logistic model trees).

2.5.4. Bayesian networks

Bayesian networks are probabilistic graphical models representing a set of random variables and their conditional dependencies via a directed acyclic graph. Each node in the graph represents a random variable. The random variable refers to a feature about which we may be unsure. Each random variable has a set of mutually exclusive and collectively comprehensive possible values. That is, exactly one of the possible values is or will be the actual value, and we are not sure about which one it is. The graph represents direct qualitative dependence relationships; the local distributions represent quantitative information about the strength of those dependencies. The graph along with the local distributions represent a joint distribution over the random variables denoted by the nodes of the graph (Neapolitan, 2004). One of the most important features of Bayesian networks is that they offer a well-designed mathematical structure for modeling complex relationships among random variables while keeping a relatively simple visualization of these relationships (Heckerman *et al.*, 1995). In this study, to select the best Bayesian network for jujubes grading, different search algorithms were evaluated.

2.6. Statistical analysis

In order to objectively investigate the performance of data mining techniques, three different statistical indicators namely, root mean squared error (RMSE), correlation coefficient (r), and correct classification rate (CCR) were considered. These indicators are mathematically calculated as follows (Mollazade *et al.*, 2012):

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(t_k - z_k)^2}{N}} \quad (2)$$

$$r = \frac{N\sum_{k=1}^{N} t_k z_k - \left(\sum_{k=1}^{N} t_k\right)\left(\sum_{k=1}^{N} z_k\right)}{\sqrt{N\left(\sum_{k=1}^{N} t_k^2\right) - \left(\sum_{k=1}^{N} t_k\right)^2}\sqrt{N\left(\sum_{k=1}^{N} z_k^2\right) - \left(\sum_{k=1}^{N} z_k\right)^2}} \quad (3)$$

$$CCR = \frac{N_{right}}{N} \quad (4)$$

Where t_k and z_k are respectively the actual and predicted value; N and N_{right} respectively belong to the total number of samples in testing set and the number of correctly classified samples.

3. Results and discussion

In order to determine the best classifier, several items were examined for each method. The results of jujube classification using the different data mining-based techniques are presented in the following sections.

3.1. Classification by artificial neural networks

Network topology is an important factor in designing artificial neural networks, because the type of topology has a significant influence on the learning rate and final classification accuracy of network. Moreover, the number of hidden layers and number of neurons in each hidden layer are main factors for designing MLP networks. These factors in turn depend on the complexity of the problem to be solved. Also, determination of the number of epochs in the learning process of the network is an important issue (Mollazade *et al.*, 2012). The number of neurons in input and output layers were fixed because they depend on independent (feature vector) and dependent (class) variables. The input layer consisted of 13 neurons (F-1, F-3, F-4, F-5, F-6, F-7, F-8, F-12, F-21, F-35, F-36, F-46, and F-53) based on feature selection operation (Table 1). Since jujubes must be graded into four classes, the output layer consisted of four neurons, each of which corresponded to one of the possible groups (G1, G2, G3, and G4). Afterwards, hidden layers were applied for developing the MLP models. In order to achieve the optimal performance for the network, several arrangements for the number of neurons in hidden layer and number of epochs were tested through trial and error procedure (Number of neurons in hidden layer varied from 2 to 20 and number of epochs varied from 100 to 1000). The best arrangement for the network was found according to the values of RMSE and r (Eqs. (2) and (3)). Results showed the hidden layer with seven neurons (i.e., 13-7-4 topology) had the lowest standard deviation (0.0067 for r and 0.0390 for RMSE) compared with the other configurations. One of the most significant points in design of artificial neural networks for online applications is proper determination of hidden layers. The lower number of neurons in hidden layer is preferred, because it results in a decrease in the networks size and consequently a decrease in the analysis time. In this study, the 13-7-4 network topology was selected as the superior architecture for jujube classification (Fig. 5). The confusion matrix of this topology using the testing data is presented in Table 2. According to the results, the classification accuracy of artificial neural network model for G1, G2, G3, and G4 classes was 100%, 100%, 97.44%, and 100%, respectively. The overall accuracy of the model was obtained as 98.61%. Confusion matrix shows the model could separate G1 and G2 grades jujubes based on the defining features successfully, but the model has the lower ability to separate G2 and G3 grades. The high accuracy of the neural network topology shows the selected features were suitable. For achieving a better performance, the shape and size features should be improved.

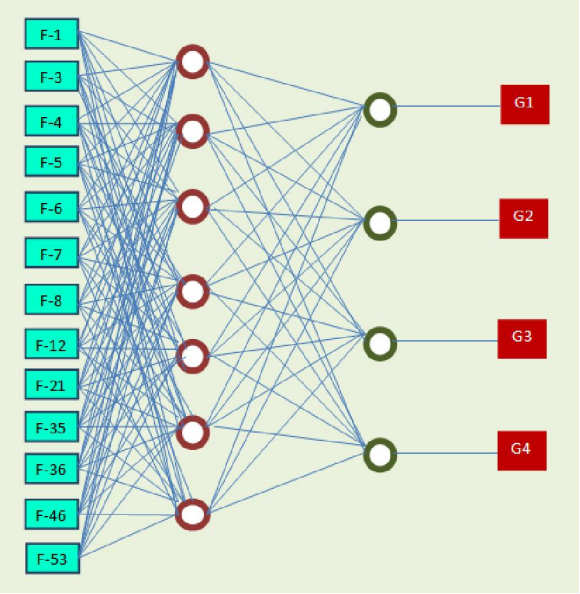

Fig. 5. The selected artificial neural network architecture with 13-7-4 topology for jujube classification.

Table 2- Confusion matrix of 13-7-4 ANN topology				
Classified as →	G1	G2	G3	G4
G1	24	0	0	0
G1	0	16	0	0
G1	0	1	17	0
G1	0	0	0	22
CCR	100	100	94.44	100

3.2. Classification by support vector machines

Kernel trick is one of the common approaches for solving nonlinear solvable problems. This technique is based on the inner product of input data, and a definition of suitable kernel function. The idea of the kernel function is to enable operations to be performed in the input space rather than the potentially high dimensional feature space. Thus, the inner product does not require to be examined in the feature space. Selection of the right kernel would improve the performance of the classifier. In this study, four common kernel functions were utilized by trial and error on the test set (Omid, 2011). The used kernel functions were namely polynomial, normalized polynomial, static kernel matrix, and universal Pearson VII (Cristianini and Shawe-Taylor, 2000). Results showed that polynomial kernel function had the highest r and lowest RMSE compared to the other functions. Hence, the polynomial kernel function was selected as the best function for jujube classification. The confusion matrix of jujube classification, using the polynomial kernel function for G1, G2, G3, and G4 classes, is given in Table 3. As shown, the classification accuracy of the support vector machines model for G1, G2, G3, and G4 classes was 100%, 93.75%, 94.44%, and 95.45%, respectively, respectively. The overall accuracy of the SVM model was equal to 95.91%.

Table 3- Confusion matrix obtained from the evaluation of support vector machines with polynomial kernel function.				
Classified as →	G1	G2	G3	G4
G1	24	0	0	0
G1	0	15	1	0
G1	0	1	17	0
G1	0	0	1	21
CCR	100	93.75	94.44	95.45

3.3. Classification by decision trees

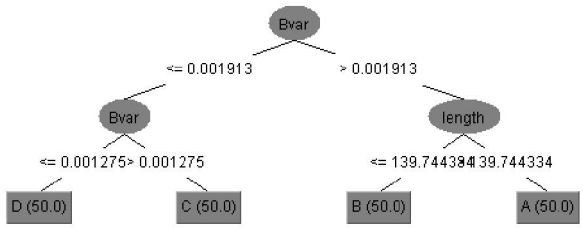

Fig. 6- Structure of J48 tree for jujube classification.

Table 4- Confusion matrix of J48 decision tree.				
Classified as →	G1	G2	G3	G4
G1	22	2	0	0
G1	0	16	0	0
G1	0	2	16	0
G1	0	0	0	22
CCR	91.67	100	88.89	100

Based on the results, J48 tree had the highest *r* (0.673) and lowest RMSE (0.1118) compared with the other trees. Therefore, this tree was selected as the best tree for jujubes classification. As shown in Fig. 6, the structure of this tree consists of 6 branches and 4 leaves. According to the confusion matrix obtained from the validation stage (Table 4), the accuracy of the system was 91.67%, 100%, 88.89%, and 100% for G1, G2, G3, and G4 classes, respectively. The overall accuracy of J48 tree for jujubes classification was 95.14%.

3.4. Classification by Bayesian Networks

When using Bayesian Networks for classification problems, the type of learning process is very important, because the accuracy of the network extremely depends on this factor. Generally, there are two learning procedures for these classifying networks; parametric learning and structural learning. The objective of structural learning is to find the best structure for the Bayesian network, which has answered with the data set and be optimum in the case of complexity. The structural learning is comprised of two method categories; limit-oriented and point-oriented. In the point-oriented method, the best network is one that has answered better with Bayesian Networks and is defined by the independent relation between nodes. In this study, five point-oriented methods namely genetic search (Generation size: 100 and population size: 10), hill climber search, K2

search, simulated annealing search (Start temperature: 10 ˚C, delta value: 0.999, and run number: 10,000) and Tabu search methods were used in the learning stage of Bayesian networks. The purpose of this procedure was to find the best learning method in which Bayesian network gives the highest accuracy in the jujube classification. In this classification technique, the Hill Climber was the best learning method for jujubes classification with the highest *r* (0.984) and the lowest RMSE (0.1118). Results of Bayesian network with simulated annealing learning for validation data set are shown in the confusion matrix (Table 5). The accuracy of network for classifying G1, G2, G3 and G4 was 91.67%, 93.75%, 100%, and 95.45%, respectively. The overall accuracy of Bayesian network was obtained as 95.22%.

Table 5- Confusion matrix obtained from the evaluation of Bayesian network with Hill Climber as search algorithm.

Classified as →	G1	G2	G3	G4
G1	22	2	0	0
G1	0	15	1	0
G1	0	0	18	0
G1	0	0	1	21
CCR	91.67	93.75	100	95.45

4. Conclusion

In this research, four different data mining techniques were used to classify jujube fruits into four qualitative grades. Comparison of validation stage of the utilized techniques indicated that MLP neural network with the 13-7-4 topology was the best classifier with an accuracy of 98.61%. After MLP network, support vector machines with polynomial kernel function, Bayesian network, and J48 tree with simulated annealing learning have higher accuracy, respectively. Comparison of confusion matrices obtained from data mining techniques showed that these techniques were very successful in separating jujubes based on the defining features. Since most of the defined size and texture features were selected in the final feature vector and contributed to the classification process, use of such features may meet the classifiers' requirements for quality grading of jujubes without the need to the color features. However, color features can be used for identifying the jujube fruits with wrinkled rinds from the ripe ones with smooth peelings. Based on the results obtained in this research, the following suggestions can be offered for the future researches:

- Fabricate a jujube sorting machine using the algorithms and techniques proposed in this paper and determine the actual classification accuracy of the system during sorting process.
- Use the optimized number of features (shape, size, texture and color) to increase the overall accuracy of the system.
- Develop the proposed algorithms here for the other external qualitative indices of jujubes.

References

1. Al Ohali, Y. 2011. Computer vision based date fruit grading system: Design and implementation. Journal of King Saud University-Computer and Information Sciences, 23: 29-36.
2. Cristianini, N. and Shawe-Taylor, J. 2000. An introduction to support vector machines and other kernel-based learning methods, Cambridge university press.
3. Gonzalez, R. C., Woods, R. E. and Eddins, S. L. 2004. Digital image processing using MATLAB, Pearson Education India.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11: 10-18.
5. Hall, M. A. 1999. Correlation-based feature selection for machine learning. The University of Waikato.
6. Heckerman, D., Mamdani, A. and Wellman, M. P. 1995. Real-world applications of Bayesian networks. Communications of the ACM, 38: 24-26.
7. Hu, B.-G., Gosine, R. G., Cao, L. and De Silva, C. W. 1998. Application of a fuzzy classification technique in computer grading of fish products. Fuzzy Systems, IEEE Transactions on, 6: 144-152.
8. Jiang, J. X. and Zhou, J. H. 2013. Dried Jujubes Online Detection Based on Machine Vision. Advanced Materials Research, 655: 673-678.
9. Karray, F. O. and De Silva, C. W. 2004. Soft computing and intelligent systems design: theory, tools, and applications, Pearson Education.
10. Khadem, E. A. F. N., E. Sharifi, M. 2013. Data Mining: Methods & Utilities. Researcher, 5: 47-59.
11. Kirkos, E., Spathis, C. and Manolopoulos, Y. 2008. Support vector machines, Decision Trees and Neural Networks for auditor selection. Journal of Computational Methods in Science and Engineering, 8: 213-224.
12. Leemans, V., Magein, H. and Destain, M.-F. 1998. Defects segmentation on 'Golden Delicious' apples by using colour machine vision. Computers and electronics in agriculture, 20: 117-130.
13. Li, X., Yuan, J., Gu, T. and Liu, X. 2009. Level detection of raisins based on image analysis and neural network. The Sixth International

Symposium on Neural Networks (ISNN 2009). pp. 343-350. Springer.
14. Mollazade, K., Ahmadi, H., Omid, M. and Alimardani, R. 2009. An intelligent model based on data mining and fuzzy logic for fault diagnosis of external gear hydraulic pumps. Insight-Non-Destructive Testing and Condition Monitoring, 51: 594-600.
15. Mollazade, K., Omid, M. and Arefi, A. 2012. Comparing data mining classifiers for grading raisins based on visual features. Computers and Electronics in Agriculture, 84: 124-131.
16. Ndehedehe, C. E. 2014. Support Vector Machine Based Kernel Types in Extraction of Urban Areas in Uyo Metropolis from Remote Sensing Multispectral Image. Researcher, 6: 105-112.
17. Neapolitan, R. E. 2004. Learning bayesian networks, Prentice Hall Upper Saddle River.
18. Omid, M. 2011. Design of an expert system for sorting pistachio nuts through decision tree and fuzzy logic classifier. Expert Systems with Applications, 38: 4339-4347.
19. Omid, M., Mahmoudi, A. and Omid, M. H. 2009. An intelligent system for sorting pistachio nut varieties. Expert Systems with Applications, 36: 11528-11535.
20. Omid, M., Mahmoudi, A. and Omid, M. H. 2010. Development of pistachio sorting system using principal component analysis (PCA) assisted artificial neural network (ANN) of impact acoustics. Expert Systems with Applications, 37: 7205-7212.
21. Qi, X.-X., Ma, B.-X. and Xiao, W.-D. 2011. On-Line Detection of Hami Big Jujubes' Size and Shape Based on Machine Vision. Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), 2011 International Conference on. pp. 96-102. IEEE.
22. Vapnik, V. 2000. The nature of statistical learning theory, springer.
23. Witten, I. H. and Frank, E. 2005. Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

5/10/2014