A Comparison Between SVM and K-NN for classification of Plant Diseases

Sarah Saadoon Jasim and Ali Adel Mahmood Al-Taei

Dept. of IT - Technical College of Management - Baghdad - Middle Technical University

sarahsaadoon9@gmail.com

Received: 23 May 2017 Accepted: 29 September 2017

Abstract

Vegetable crops differ in size, shape, and color and which its suffer from this many leaf batches according to a particular reason. As a result of the plant, pathogens happen for Leaf batches. In agriculture whole fructification, it is essential to learn the origin of plant disease bundles early to be prepared for suitable timing control. In this regard, uses Support Vector Machine (SVM) and K- Nearest Neighbor to classify the plant's symptoms according to their appropriate classifications. These typesare (YS) Yellow Spotted class, (WS) White Spottedclass, (RS) Red Spotted class, and (D) tarnishedclass. Results obtained using SVM algorithm was compared with results obtained by a K-NN algorithm. Specifically, the overall accuracy of SVM model is about 88.17% and 85.61% for the k -NN model (with k = 1).

Keywords: Classification, Support Vector Machine (SVM), k- Nearest Neighbor (k-NN).

مقارنة بين طريقتي SVM و K-NN لتصنيف امراض النبات

سارة سعدون جاسم و علي عادل محمود

قسم تقنيات ألمعلوماتية – ألكلية التقنية ألأدارية – بغداد – ألجامعة ألتقنية ألوسطى

ألخلاصة

تختلف محاصيل الخضروات من حيث الشكل والحجم واللون كما وتعاني هذه المحاصيل من بعض الامراض التي تظهر على شكل بقع على الاوراق. في علم الزراعة من المهم ان نعلم اسباب ومصادر الامراض التي تصيب المحاصيل من اجل التهيؤ للوقاية منها قبل حدوثها او علاجها باقل الخسائر، ولتحقيق هذا الهدف فقد تم استخدام طريقتي SVM و K-NN من اجل تصنيف اوراق النبات وكشف ما اذا كانت مصابة ام لا. هنالك اربعة تصنيفات اساسية: البقع الصفراء، البقع البيضاء، البقع الحمراء والتشوه الحاصل في الورقة. اظهرت النتائج ان استخدام طريقة SVM افضل واسرع في تمييز امراض النبات.

الكلمات المفتاحية: التصنيف، طريقة الجار الاقرب (K-NN)، طريقة الة المتجه الداعم (SVM).

Introduction

There are many types of diseases that vegetable crops suffer from, and one of those diseases is leaf batches. Plant pathogens which related to fungi, bacteria and virus diseases, insect feeding which related to sucking insect pests and plant nutrition which related to micro elements are the causes of leaf batches. According to the different causes of this disease, leaf batches have some characteristics such as (color, shape, and size). It is playing an important role to determine the cause of the disease; Sometimes those characteristics cause confusion when they used in diagnosing leaf batches because of the similarities in batch size, shape, and color but, the only expert can identify them. In agricultural production, it is very important to discover the disease in its early time or at the beginning of its first stage to lessen the harm, minimize the production expenditure and to improve the returns.That the essential symptom which refers to the existence of this disease is the leaf spots; therefore, spots can be used to determine the cause of this disease. Two steps are needed for determining the cause of leaf batches, initially is to extract

the feature of the leaf spots such as color, size, and shape, and second is using a classifier to learn from those characteristics and contrast between them [1]. In this paper, we used (SVM and K-NN), as a tool to classify the plant symptoms according to their appropriate categories, which are (YS) the yellow spotted, (WS) white spotted, (RS) red spotted, (N) Normal and (D) discolored spotted. The results obtained by the SVM algorithm were compared with the results of the K-NN algorithm.

Literature review

In agricultural applications, researchers using some techniques such as image processing and pattern recognitions for many years. The process of diseases detection is performed by grade and sortthe vegetables and fruits which have been packed in packing houses. For example, The procedure of those techniques is the same and it can be implemented by specific steps: firstly is acquiring an image by capturing pictures for the real state environment for the field using any kind of camera. Secondly, utilizing one of the most dominant feature extraction methods, to extract a meaning full features of the captured image in the future analysis. Thirdly, implement a classification technique for the extracted features such as parametric or non-parametric techniques or artificial neural network to classify the collected image and come out with study objectives. In any machine learning system which implement image processing techniques the classification process plays the key role among other steps due to the complexity. In addition, the direct affect for the study results. Mohammed J. Islam, Q. M. Jonathan Wu et al. (2010) used two methods Bayesian theory the N-Bayesclassifier and K- Nearest Neighbor classifier. After the implementation and applied these methods to database "Credit card approval" application. In K-Nearest Neighbor method, they perform the classification for the same dataset by selecting different values of k, when the value of k is equal to (5) we outcome with the highest classification rate and the error rate was at (9.45%) [2]. Savita N. Ghaiwat, Parul Arora (2014) uses several different techniques such as K-NNClassifier, Probabilistic NN, GA, SVM and PCA, ANN, Fuzzy logic to classify plant leaf disease. This study has resulted the K-NN method represent the simplest one a many other over in terms of estimating the right class [3]. Jagadeesh D.Pujari, Rajesh Yakkundimath et al. (2016) presents a study to reduce the features of images to distinguish and classification of plant diseases. Certain algorithms have been

developed to extract some features of color and texture and which were used to train SVM and ANN classifiers. The study proved that SVM method has presented the best performance in terms of accuracy estimated (92.17%)and this ratio is better than the ANN classifier where accuracy value estimated (87.48%) [4]. Jasmeet Kaur et al. (2016) used BP, PCA combined with SVM to increase the detection accuracy of the diseased plant leaves. The plant diseases can be identified at the initial stage itself and control can be obtained. Although the algorithm discrimination was optimized to get better results when we have reduced the dimensions of the features of the data, so that led to less accurate. The SATURATION value of infected leaf will be calculated which helps to increase detecting accuracy. The SVM shows the portion that how much part is infected, which is 5.54 % [5].

Lakhvir Kaur1, Dr. Vijay Laxmi (2016) used KNN model to classify leaf venation morphological feature that was classified them into four different species. Study reached across the method performance was (96.53%)and (91%) ofaccuracy for training and testing respectvely to classify the images of leaves plants [6].

2- Classification

Supervised and unsupervised learning are two types of methods of Classification. In the method of unsupervisedlearning, there is an outstanding method which is clustering. Different types of clustering analysis methods are available such as: viz. DB-SCAN (Ester et al., 1997), K-Means, XMeans (Pelleg and Moore, 2000) or SVM-Clustering (Ben-Hur et al., 2001). In neural network, the learning algorithm can be either supervised or unsupervised. However, a supervised type of learning algorithms is applicable when the desired output is already known. On the other hand, unsupervised type of learning algorithms is employed if no target output is available [7].

2.1 Support Vector Machines

To perform nonlinear class boundaries, Support Vector Machines uses linear models for this purpose. The input space is transformed into a new space (F feature space) using a nonlinear mapping. After that to represent a nonlinear decision boundary in the original space the linear

model that has been constructed from the new space are used. Figure (1) showes the SVM classification concept (as adopted from [3]) .

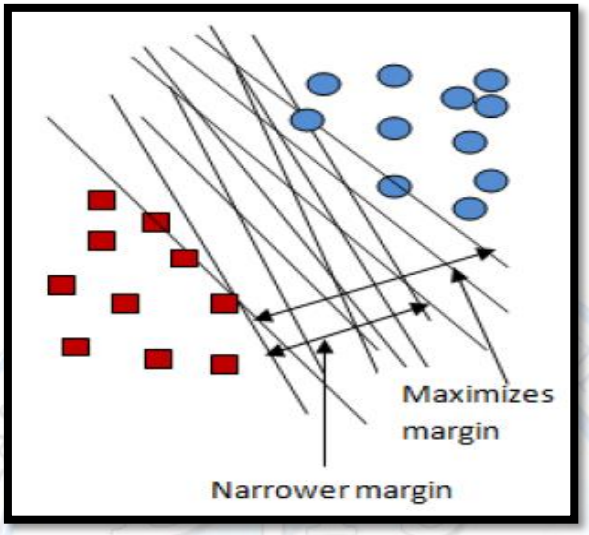

Figure 1: SVM classifier

There is another important component in the SVM approach which is known as highest margin hyperplane. The highest margin hyperplane can be defined as the greatest separation between two classes dataset, which are linearly separable, and if we have many maximum margin hyperplane that separates the classes. The maximum edge can refer to that as far away as possible from two convex objects, each of which is consists by connecting to class instances. Figure (2) showes the highest margin hyprplane approach (as adopted from [8]).

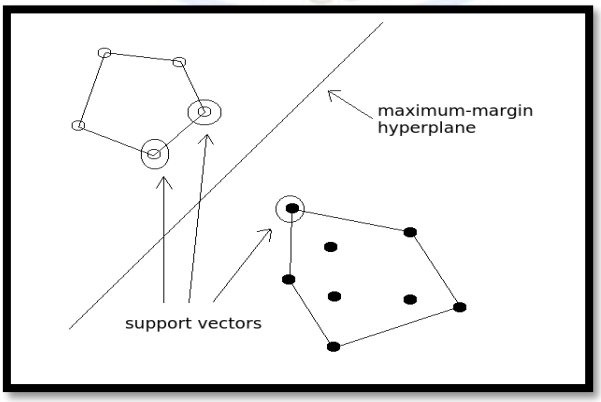

Figure 2: Maximum margin hyperplane

There are two types of training instances; one is called support vectors; these types of instances vary close to the maximum margin hyperplane. Each class may contain one or many support vectors. A group support vectors can individually define highest margin hyperplane for the learning problem. The rest of instances were meaningless and ignored regardless changing the positions as well as the orientation of the hyperplane. Basically, in two-attribute case, the equation that is used to represent the hyperplane separating two classes can be formulated as ($F(x) = w_0 + w_{a1} + w_{a2}$), where a1, and a2 are the attributes values, and w_0, w_1, w_2 are weights. The highest margin hyperplane equation can be written as ($F(x) = b + \sum_{i=1}^{1} a_i\, y_i \langle x_i\, .\, x\rangle$), depending on the support vectors terms. Where i is the support vector;

$y_i$represent the training value of the class: either 1 or -1;

x_i (i^{th} support vector);

x (a test state vector);

$\langle x_i . x\rangle$represent the dot product of x and x_i;and b and $\alpha_i$are parameters just as the weights w0, w1, w2 that determine the hyperplane. Finding support vectors and the maximum margin hyperplane (i.e., b and α's) that going to a standard class of optimization problem famous as quadratic constraint optimization. SVM has many advantages; it generates an accurate classifier, less overfitting, and robust to noise. However, it has disadvantages too. It runs slow because SVM is a binary classifier, it will do (a multi-class classification, pair-wise classifications) such that may be utilized as single class faces the rest of classes. And therefore, lead to computationally expensive [3].

2.2 K-Nearest Neighbor

It represents one of the simplest classification techniques that used under machine learning context. For instance, it is based on the perform of classification by specifying the nearest neighbors to target, and fixing the class of query to be done benefit of these neighbors. In the K-NN classification, it calculates the least distance between a particular point and other points. K-NN classifier in which the nearest neighbor doesn't contain any training process. A large

number of training examples is not usable in this case, and it is not strong to noisy data. The distance is calculated between (test and training) samples for classification of a plant leaf. Such method brings up similar measurements according to the class of test samples and classifies the sample based on the highest number of votes from the k-neighbors. For instance, the sample is also set to the class most common among its k-nearest neighbors. K is a small positive integer, if k = 1, so the sample is set to the class of its nearest neighbor. In two classes (binary) of classification problems, choose k an odd number that is useful to avoids balanced votes.

One of the advantages of (K-NN) algorithm is its easy implementation and present significant performance especially when the features are perfectly selected firstly. Secondly, K-NN Classifier is worked more on basic recognition problems. The main drawback of the KNN method is firstly low-speed performance, i.e. simply it uses the same training data for classification, and it is not learning anything from the training data. Secondly, the method calculates all distances among the nighbors which cause heavey processing resources when the dataset is huge. Thirdly, when the data under process have some noise, the method performance is affected negatively, which means this method is not robust enough with noisy data. Fourthly, the most drawback of this method is that the high sensitivity of the presence of the meaningless values [3].

3- Proposed Methodology

In this study, we present a method that focuses on detectingthe disease thatappearss on plants and affecting them early enough to handle suchsituationss. The real world that consist of the tasks similar to image acquiring feature selection, pre-processing of images, formulating methods to be applied, and developing the architecture of the classification model for the computer vision system. The Figure (3) illustrates the main faces of the proposed method.

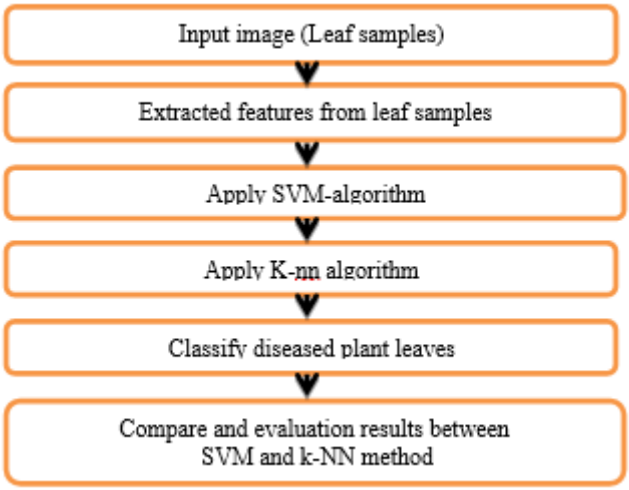

Figure 3: Block diagram for proposed system

3.1 Image data set

In our study, the diseases that have been founded on different types of plants such as fungal, bacterial, viral, nematodes, and normal (not affected) utilized to be implemented in the study processes, and so that appoints five classes. The study considers 1868 sample images for the method processes. However, the sample dataset of our study were selected from (Central Lab of Agricultural and expert systems (CLAES), Cairo, Egypt. The first symptoms and signs of the disease were noticed on leaves, the methods for the identification and classification of plant diseases affecting agricultural leaves and limiting their development. Figure (4) shows the classification tree.

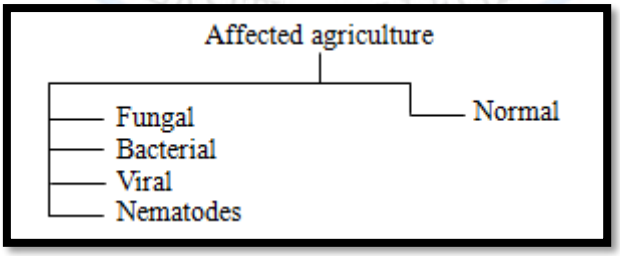

Figure 4: Classification tree

3.2 Feature extraction

In this paper, feature extraction involved the features that have been selected from segmented image, which is used in classification to discover the spot category. First, the features depend

on the characteristics of the colors of the captured image, for example, the gray-level mean, red, green, and blue colors for the captured spots, while the second features are related to the morphological characteristics of the spots [1].

•*Major and Minor axes length of a spot are the length of the principal axes:*

The diameter of a spot is $d = \sqrt{\frac{4\times area}{\pi}}$ .....(1)

• *Eccentricity or circularity ratio (CR), its value are located between the spot whose circularity ratio is zero (actually a circle), while the spot whose circularity ratio is one is a line. The circularity ratio is computed as*$CR = \frac{\sqrt{(x)^2-(y)^2}}{x}$ ..... (2), such that (major=x and minor=y).

• *Compactness Measure,* also called solidity ratio, has a value between [0,1]. Essentially, a fully compacted spot is any captured spot that has fixed value of one, however, it can be evaluated via the ratio of the spot area to the convex hull (i.e. the form is computedas)

$\text{Ratio} = \frac{\text{spot area}}{\text{convex hull}}$ ..... (3)

• *Extent Measure,* also called rectangularity ratio, has a value between [0, 1], when this equal ratio one, then the spot shape will be typically rectangle, extent measure is computed as

$EM = \frac{spot\ area}{boundingbo\ xarea}$ ..... (4)

• *Euler's Number Measure*: This measure describes a simple topologically invariant property of the spot. It is calculated by subtracting the number of holes from the object regions.

• *Orientation Measure*: Is the angle in degrees between the, and the major axis length of the spot. axis-x

Building a classifier which is capable of classifying leaves symptoms is our main goal. We have distributed disorders in each class as follows: Te first class (*YS*) contains eleven disorders, which are Downey, Leaf Blight, Leaf Spot, High Temp, Jassid, Magnesium Def, Potassium Def, Salt Injury, Scab, Spider, and Zinc Def. The second class(*WS)*contains nine disorders, which are Aphids, Magnesium Def., Leaf Miner, Powdery, Manganese Def, Slat Injury, White Fly, Leaf Spots, and Tobacco Virus. The third class(*RS)*contains nine disorders, which are Leaf Blight, Downey, Anthracnose, Leaf Spot, Gummy Stem Blight, Pesticide Injury, Phosphorus

Def, Spider, and Toxicity. The fourth class(*D)*contains nine disorders, which are Downey, Mosaic, Iron Def., Manganese Def., Nitrogen Def., Potassium, Pesticide Injury, Salt injury, and Trips. Images are collected for each class as follows, for each of the *WS*, YS, and RS class 20 images, for the D class 32 images, and for the normal class 25 images. The analysis of images shows that there is a relationship between those classes [1].

- 44%-55% overlaps the (D) class with other classes.
- 33%-55% overlaps the (YS) class with other classes.
- 33%-55% overlaps the (RS) class with other classes.
- 33%-55% overlaps the (WS) class with other classes.

Table 1: Total number of Patches for each class for the training set.

Class	Number of images	Total number of patches
RS	20	259
WS	20	514
YS	20	185
D	32	337
N	25	573
Total	117	1868

Results

Two methods were applied: SVM and k-NN to explore the best performance model. To achieve best and reliable test results, cross validation task was performed. Tables 2 and 3 illustrate the confusion matrix of SVM and k-NN, respectively.

Table 2: Confusion matrix of SVM

	A	B	C	D	E
A	170	10	1	11	0
B	20	164	28	27	17
C	7	38	387	15	12
D	2	7	7	52	2
E	10	68	41	30	515

Table 3: Confusion matrix of k-NN

	A	B	C	D	E
A	166	6	7	14	3
B	13	164	25	16	40
C	8	35	387	13	33
D	10	20	6	70	11
E	12	62	39	22	459

Class precision and recall parameters for each class were calculated and compared. Table 4 shows the values of class precision and class recall parameters for each single class of data, in both tested models (SVM and k-NN).

Table 4: Performance of models

	SVM		k-NN	
	Class precision	Class recall	Class precision	Class recall
A	88.54%	81.34%	84.69%	79.43%
B	64.06%	57.14%	63.57%	57.14%
C	84.31%	83.41%	81.30%	83.41%
D	74.29%	38.52%	59.83%	51.85%
E	77.56%	94.32%	77.27%	84.07%
Avg. Error	11.83%		14.39%	
Avg. Accuracy	88.17%		85.61%	

Conclusions

Plant disease recognition is an important growing area of research. In this paper, two methods were applied: SVM and k-NN on data of plant disease that appear on an infected leaf. The stratified cross validation technique was performed on both models, the best accuracy result found was when k=10 folds. However, the overall accuracy of SVM model is about 88.17% and 85.61% for the k -NN model (k=1) with classification error of 11.83% and 14.39% for SVM and k-NN, respectively. These results over perform results in [1]. However, the SVM model in this work is close in performance to the model proposed in [9] which takes too much calculations and time (about 585 minutes), while in our model time taken is less than one minute.

References

1. Mohammed S. and Khaled S. S. Z., "Support Vector Machine Vs an Optimized Neural Network for Diagnosing Plant Diseases," Cairo University, ICENCO 2006.
2. Islam M. J., Q. M. Jonathan Wu., Ahmadi M., Sid-Ahmed M. A., "Investigating the Performance of Naïve- Bayes Classifiers and K- Nearest Neighbor Classifiers," Journal of Convergence Information Technology Volume 5, Number 2, April 2010.
3. Ghaiwat S. N., Arora P., "Detection and Classification of Plant Leaf Diseases Using Image Processing Techniques: A Review," International Journal of Recent Advances in Engineering & Technology (IJRAET) ISSN (Online): 2347 - 2812, Volume-2, Issue - 3, 2014.
4. D. Pujari J., Yakkundimath R., and Abdul Munaf. S. B., "SVM and ANN Based Classification of Plant Diseases Using Feature Reduction Technique," International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 3, Nº7- 6 -DOI: 10.9781/ijimai.2016.371.
5. Kaur J., and Ramanpreet Kaur Er., "Plant Disease Detection using SVM Algorithm and Neural Network Approach," International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, Vol. 4, Issue 6, June 2016.
6. Kaur L. and Laxmi V., "Detection of Unhealthy Region of plant leaves using Neural Network," International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137 Volume – 01, Issue – 05, August – 2016.
7. Rumpf T., Mahlein A.-K., Steiner U., Oerke E.-C., Dehne H.-W., and Plümer L., "Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance", journal homepage: www.elsevier.com/locate/compag, Computers, and Electronics in Agriculture 74 (2010) 91–99.
8. Internet Survey, http://www.kdnuggets.com/2016/09/support-vector-machines-concise-technical-overview.html, Accessed at 29/8/2017.
9. Mohammed S. and Mohammed El-Beltagy, "Optimizing Neural Networks Architecture and Parameters Using Genetic Algorithms for diagnosing Plant Diseases", Proceeding of International Computer Engineering Conference, IEEE, Egypt 2006.