International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Challenge Faced in K-Means Clustering for the Choice of Automatic K-Value for Segmentation of Black Sigatoka Disease in Banana Leaves

Srivalli Devi S., A. Geetha

Abstract: ***Clustering is defined as grouping similar items . The three types of machine learning techniques are supervised, unsupervised and semi-supervised. In unsupervised technique, there are no class labels given to the input data. Clustering is a type of unsupervised learning technique. Recently clustering is applied in many fields such as medicine, agriculture, biology, computers, finance and robotics. Black sigatoka is a bacterial disease occurring commonly in banana plants .The research currently focuses on segmenting the disease area from non-diseased area.The segmentation class training is done via Trainable Weka Segmentation and we also do segmentation using k-means algorithm. In this paper we propose a novel approach for extraction of the black sigatoka diseased area on banana leaves from images using pixel color values and grouping them into their respective clusters accordingly. This is a segmentation cum clustering algorithm. The novel approach has been proposed to overcome the shortfall of k-means clustering when segmenting using automatic value selection for k-means by using silhouette values.Using this novel approach its easy to cluster and segment at the same time. The segmented image from this algorithm can be used in disease classification tasks.***

Keywords: Black **sigatoka,K-means clustering, Trainable Weka segmentation, silhouette score, pixel color values.**

I. INTRODUCTION

Agriculture is the major source of food for humans. But when disease occurs in the cultivated plants, these plants tends to be a waste for consuming purposes. Hence there comes the role of identification of disease during even early stages using modern technology. Farmers can access the disease detection software directly using their mobile phone by taking a picture and upload it in the software. Then the diseased areas can be segmented and classified using data mining techniques K-means clustering algorithm is the popular algorithm that is used for unsupervised image segmentation. For image mining , there are different kinds of steps such as image acquisition, image processing , classification, color processing, clustering and feature extraction[1].

Revised Manuscript Received on January 30, 2020.
* Correspondence Author
Srivalli Devi S.*, PG & Research, Department of Computer Science, Chikkanna Government Arts college, Tirupur (Tamil Nadu) India. E-mail: ssrivallidevi@gmail.com
Dr. A. Geetha, PG & Research, Department of Computer Science, Chikkanna Government Arts college, Tirupur (Tamil Nadu) India. E-mail: gee_sam@yahoo.com

II. LITERATURE REVIEW

The k-means algorithm is widely used algorithm for segmentation using clustering techniques.
K-means was used as primary segmentation algorithm for segmenting medical images along with improved watershed algorithm was used[2].Using k-means has reduced the over-segmentation in MRI images.In [3] authors used a method for k-means in which the mean square error is reduced in generating the cluster center. Execution time was not affected too. They determined that dense dataset achieves higher accuracy. In [4] authors suggested for brain tumor segmentation to use k-means and fuzzy C-means algorithm for calculating the area of the tumor in the brain. They used MRI images. A novel method for discovering the initial centroid was proposed in [5] and authors of that paper states a better way for data point assignment with time complexity lesser than that of others for suitable clusters.Initial centroid was computed using subtractive clustering technique and it was then applied to k-means algorithm for segmentation and partial contrast stretching used for image quality improvisation and median filter for improvisation of segmented image. This idea was used in [6] for medical image segmentation.In [7] the authors had masked the green pixels and used colour co-occurrence method and suggested that for the accuracy improvement neural networks may be used.In [8] authors had used the clustering technique k-means clustering and also they used it with neural networks for leave disease detection.In [9] authors had used co-occurrence matrix and k-means for segmentation of texture. They also suggest that k-means , Bayes classifier and principal component analysis can be used to classify various plant diseases.Genetic algorithms with k-means clustering technique was used in [10]. The green pixels are masked and the diseased portion is thus obtained and they pass the clusters from k-means to genetic algorithm for better accuracy in diagnosing the plant leaf diseases.

III. PROPOSED METHODOLOGY

In the existing systems, k-means is used by giving the default value for k done by predetermining the number of initial clusters. But this method of k-value default assignment deducing the number of clusters required beforehand may sometimes lead to unexpected results where there may be more actual number of clusters than predetermined. In this paper we discuss on using automated k value selection and the challenge faced when we take the automatic assignment of k-value in k-means.

A.Dataset used

The dataset used here contains 42 banana leaves images[11] that are affected with black sigatoka disease. Black sigatoka is a fungal disease commonly affecting the banana and plantain plants. The pathogen that causes this disease is Mycosphaerella fijiensis.

B. K-means

The automatic selection is done by calculating silhouette score using Euclidean distance and Manhattan distance. The average of these are computed and considered for picking up the automatic k-value for segmentation. It is an unsupervised clustering based on centroid based approach. The algorithm is as follows. The k value for k-means algorithm is estimated using silhouette value calculated with Euclidean distance and Manhattan distance and the average of these two values are calculated for each image and for each image separate k value based on higher silhouette value is selected and this k value is passed to the k-means algorithm for performing clustering. The image and the silhouette score with their k-value are given in the table. The k-means clustering algorithm is used from [12].

C. Silhouette score

For a data instance, silhouette is a measure of how closely is matched within its cluster or loosely it is matched to the neighbouring cluster [13]. In Table-I the silhouette scores for different banana leaves affected with black sigatoka disease is calculated for selecting the k value to provide to k-means algorithm. Then k-means algorithm is used for clustering.

D. Trainable Weka Segmentation in ImageJ

Trainable Weka Segmentation is distributed as open-source software as part of the Fiji image processing distribution of ImageJ at [14].Trainable Weka Segmentation used for pixel classification. It can be applied for boundary detection, semantics segmentation, or object detection and localization. It's a GUI based application in machine-learning. In Trainable Weka Segmentation plugin, the fast random classifier is used for segmentation[15]. The segmented images as shown in the Table2 segments the affected and non-affected portions. But the segmentation when compared with k-means segmented images, trainable weka segmented images is not up to the mark for our problem to find segmented region of the black sigatoka disease.

IV. DEV ALGORITHM

We are suggesting a novel approach which we named as Dev algorithm which does segmentation and clustering together in a single algorithm. In this novel approach, both segmentation and clustering is done one after another.

The black sigatoka disease affected banana plant leaf image is taken. We use Python for the programming part. Then one by one red, green and blue component (RGB value) of each pixel is read from the image. Then we pre-compute the diseased region pixel range RGB value. Now comparing the RGB values of each pixel with the RGB value of the diseased RGB component, two of the following things happen. If those values fall within the diseased range color component value, then the corresponding position of the pixel is grouped in the cluster of diseases. Otherwise in non-disease cluster along with masking of those pixels. Thus the pixels are grouped into two different clusters and also the diseased area is segmented from non-diseased are by masking those pixels.

Dev algorithm steps are as follows:

Input:

Let, C_n be the non-disease cluster

C_d be the disease cluster

I be the image inputted

P_c be the current pixel

R_{Pc} be the red channel of the P_c

G_{Pc} be the green channel of the P_c

B_{Pc} be the blue channel of the P_c

R_{range_upper} be the red channel upper limit range

G_{range_upper} be the green channel upper limit range

B_{range_upper} be the blue channel upper limit range

R_{range_lower} be the red channel lower limit range

G_{range_lower} be the green channel lower limit range

B_{range_lower} be the blue channel lower limit range

Algorithm:

for each P_c in I :

if[(R_{Pc} within range of R_{range_lower} to R_{range_upper}) and
(G_{Pc} within range of G_{range_lower} to G_{range_upper}) and
(B_{Pc} within range of B_{range_lower} to B_{range_upper})]
Assign P_c to C_d

Else

Assign P_c to C_n

The algorithm can be defined in a step-by-step way as follows:

1. Each pixel from the image is read and taken one-by-one
2. For each pixel, red channel,green channel and blue channel values are obtained.
3. Now after obtaining those values as said in step 2, the value of the red,blue and green channels are checked if they fall in the range of the colour value of the diseased region. This range has to set manually depending upon the disease.
4. If the range and the pixel values coincide then the pixel is assigned to the disease cluster otherwise its assigned to the non-diseased cluster.
5. So we can retrieve the clusters and the segmented values at the same time.

V. RESULT

The following images Fig.1,Fig.2,Fig.3,Fig.4 and Fig.5 are obtained from running the algorithm on very few sample of banana leaf images. The two tables Table-I and Table-II show the values for silhouette scores and the respective segmented images obtained for the selected silhouette values.

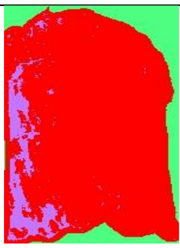
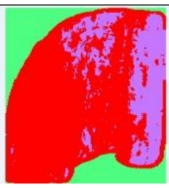
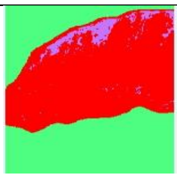
Table- I: Silhouette values corresponding image segmentation table

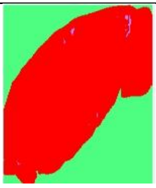
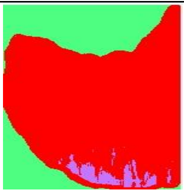
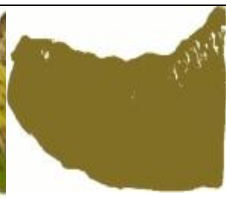
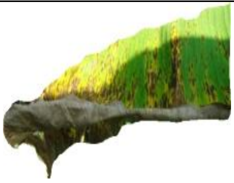
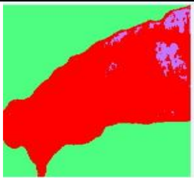
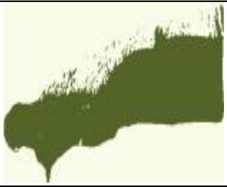
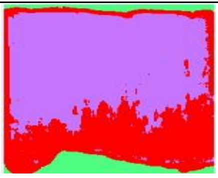
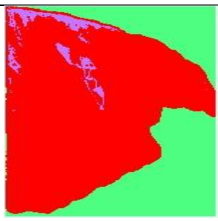
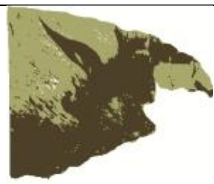
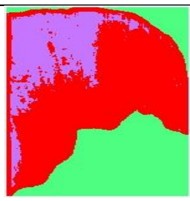
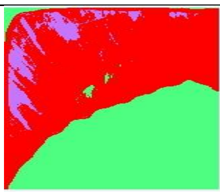
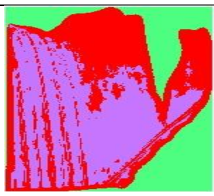
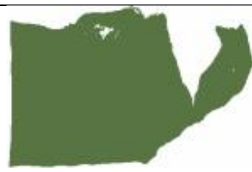
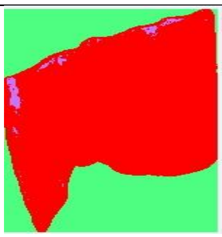
file_no	kvalue	Silhouette score using Euclidean distance	Silhouette score using Manhattan distance	(Silhouette score) Average distance =(Euclidean+Manhattan)/2	Selected<marked as 1 , otherwise>
1	**2**	**0.79171308**	**0.80579**	**0.798752073**	**1**
1	3	0.57331531	0.59716	0.585236588	0
1	4	0.51860176	0.52974	0.524170066	0
1	5	0.52366997	0.53145	0.52756058	0
1	6	0.47309512	0.47621	0.474652597	0
1	7	0.47151469	0.4765	0.47400944	0
1	8	0.44839676	0.45241	0.450403493	0
1	9	0.43859566	0.43879	0.438692545	0
2	**2**	**0.83626623**	**0.84727**	**0.841766673**	**1**
2	3	0.57378732	0.58988	0.581831934	0
2	4	0.51927553	0.5316	0.525439736	0
2	5	0.4809272	0.49111	0.486016548	0
2	6	0.45351199	0.45435	0.453929592	0
2	7	0.4625149	0.46353	0.463023011	0
2	8	0.46052831	0.46277	0.461649292	0
2	9	0.46431341	0.4659	0.465107156	0
3	**2**	**0.84484814**	**0.85277**	**0.848809382**	**1**
3	3	0.74848693	0.74686	0.747672539	0
3	4	0.69262657	0.69455	0.693587916	0
3	5	0.68673909	0.6874	0.687069733	0
3	6	0.692076	0.69434	0.693209858	0
3	7	0.67807245	0.68055	0.679311085	0
3	8	0.6671776	0.67023	0.668705565	0
3	9	0.66462369	0.66855	0.666585954	0
4	**2**	**0.74479308**	**0.75811**	**0.751453855**	**1**
4	3	0.58059141	0.59625	0.588419481	0
4	4	0.53632113	0.54925	0.542786646	0
4	5	0.49129066	0.49956	0.495423178	0
4	6	0.47131143	0.47516	0.473236639	0
4	7	0.460194	0.46308	0.461636931	0
4	8	0.4711404	0.47589	0.473516396	0
4	9	0.4585395	0.45656	0.457550307	0
5	**2**	**0.82619709**	**0.83786**	**0.832029526**	**1**
5	3	0.72601442	0.73377	0.729890897	0
5	4	0.71127792	0.71234	0.71180955	0
5	5	0.69586546	0.70112	0.698492244	0
5	6	0.6923875	0.69609	0.694241206	0
5	7	0.6855038	0.69045	0.687979003	0
5	8	0.66750103	0.6719	0.669698871	0
5	9	0.64302923	0.6486	0.645814789	0
6	**2**	**0.80922329**	**0.81873**	**0.813976372**	**1**

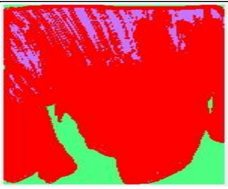
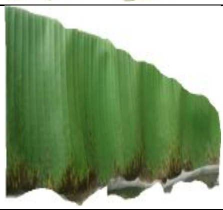
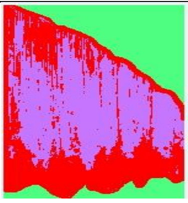
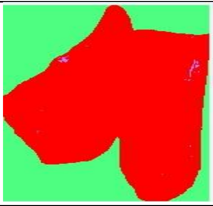
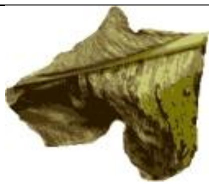
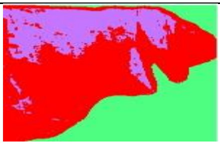
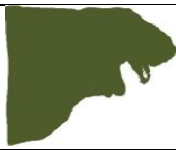
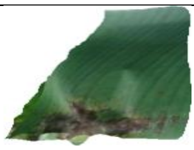
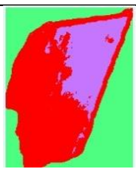
6	3	0.62208326	0.63108	0.626583372	0
6	4	0.54497835	0.55419	0.549584987	0
6	5	0.57097455	0.56888	0.569928186	0
6	6	0.57982763	0.57025	0.575037173	0
6	7	0.57524854	0.57323	0.574237665	0
6	8	0.57480213	0.57199	0.573397714	0
6	9	0.5585139	0.56134	0.55992875	0
7	2	0.781434	0.8081	0.794768978	0
7	**3**	**0.81143416**	**0.81065**	**0.811041989**	**1**
7	4	0.75505464	0.76145	0.758252355	0
7	5	0.7250953	0.71428	0.719686167	0
7	6	0.72917583	0.71756	0.723366468	0
7	7	0.71945083	0.7109	0.715176173	0
7	8	0.72248612	0.71528	0.718882365	0
7	9	0.71839306	0.71004	0.714217742	0
8	**2**	**0.91018343**	**0.91905**	**0.914614605**	**1**
8	3	0.48534419	0.49486	0.490102263	0
8	4	0.54838493	0.54626	0.547324179	0
8	5	0.55531746	0.55243	0.553874607	0
8	6	0.54672592	0.54731	0.547019785	0
8	7	0.49328732	0.50781	0.500548005	0
8	8	0.48800339	0.49684	0.492423531	0
8	9	0.49162565	0.50022	0.4959211	0
9	2	0.71865081	0.72931	0.723981676	0
9	**3**	**0.75616888**	**0.76402**	**0.76009244**	**1**
9	4	0.68783763	0.6975	0.69266679	0
9	5	0.65981659	0.67019	0.665003179	0
9	6	0.63667892	0.64558	0.641128364	0
9	7	0.6116141	0.61911	0.615363901	0
9	8	0.59575481	0.60046	0.598109174	0
9	9	0.58066559	0.58306	0.581862886	0
10	**2**	**0.83274447**	**0.84478**	**0.838759792**	**1**
10	3	0.6531277	0.66949	0.661308548	0
10	4	0.62175476	0.61461	0.618181186	0
10	5	0.6012237	0.59635	0.598786471	0
10	6	0.62312177	0.61384	0.61848138	0
10	7	0.60129441	0.59344	0.597365347	0
10	8	0.61205075	0.61074	0.611393372	0
10	9	0.60543753	0.60837	0.606904284	0
11	**2**	**0.81160099**	**0.81664**	**0.814120229**	**1**
11	3	0.77097308	0.78026	0.775616992	0
11	4	0.69706497	0.70802	0.702543392	0
11	5	0.71036445	0.72243	0.716395373	0
11	6	0.69406961	0.70619	0.700130791	0
11	7	0.68709314	0.69857	0.692833377	0

11	8	0.67025316	0.67947	0.674860311	0
11	9	0.66369706	0.67266	0.668177808	0
12	**2**	**0.87800681**	**0.88602**	**0.882014098**	**1**
12	3	0.58789787	0.6048	0.596346845	0
12	4	0.60860233	0.61016	0.609383598	0
12	5	0.5865889	0.58947	0.588029019	0
12	6	0.5700473	0.57295	0.571496199	0
12	7	0.59414999	0.59777	0.59595888	0
12	8	0.5783861	0.58275	0.580566102	0
12	9	0.5537557	0.55694	0.555349913	0
13	**2**	**0.81646821**	**0.82067**	**0.818568547**	**1**
13	3	0.71170363	0.7202	0.715950918	0
13	4	0.68520702	0.69537	0.690290923	0
13	5	0.66771758	0.67929	0.67350572	0
13	6	0.6503633	0.65995	0.655154522	0
13	7	0.63400227	0.64274	0.638370663	0
13	8	0.61925869	0.62678	0.623019236	0
13	9	0.60481544	0.6113	0.608058232	0
14	**2**	**0.69273805**	**0.70936**	**0.701049993**	**1**
14	3	0.55635729	0.5744	0.565379713	0
14	4	0.50366647	0.51396	0.508815172	0
14	5	0.51234526	0.51262	0.512481738	0
14	6	0.47063488	0.46818	0.46940869	0
14	7	0.49302695	0.49093	0.491979098	0
14	8	0.48118575	0.47967	0.480427033	0
14	9	0.49188183	0.49363	0.492754788	0
15	**2**	**0.87191793**	**0.87756**	**0.874737165**	**1**
15	3	0.62475899	0.64143	0.633092554	0
15	4	0.59092711	0.60132	0.596121942	0
15	5	0.5951703	0.60137	0.598272219	0
15	6	0.5709856	0.57748	0.574232736	0
15	7	0.55355616	0.55834	0.555950119	0
15	8	0.54778109	0.552	0.549890949	0
15	9	0.56353181	0.57058	0.567057305	0
16	**2**	**0.68353928**	**0.69473**	**0.689134937**	**1**
16	3	0.58513753	0.60299	0.594062999	0
16	4	0.57136939	0.58394	0.577654986	0
16	5	0.55303107	0.54716	0.550094409	0
16	6	0.53576225	0.52702	0.531391714	0
16	7	0.55807019	0.54484	0.55145556	0
16	8	0.57876089	0.56807	0.573416869	0
16	9	0.55098279	0.54797	0.549474937	0
17	**2**	**0.79960849**	**0.80631**	**0.802959142**	**1**
17	3	0.72770182	0.73793	0.732813524	0
17	4	0.69649125	0.70431	0.700398671	0
17	5	0.67367522	0.67749	0.67558062	0
17	6	0.65159753	0.65726	0.654427928	0
17	7	0.66117113	0.66645	0.663809392	0
17	8	0.6452303	0.65007	0.647648011	0

17	9	0.6509536	0.65681	0.653881052	0
18	**2**	**0.89560578**	**0.90475**	**0.900175566**	**1**
18	3	0.63226036	0.64866	0.640459091	0
18	4	0.60650179	0.5975	0.601999519	0
18	5	0.60482538	0.59943	0.602130177	0
18	6	0.62567831	0.61997	0.622825843	0
18	7	0.62228715	0.61807	0.620180477	0
18	8	0.61662325	0.6208	0.618712352	0
18	9	0.58380276	0.58698	0.585389073	0
19	**2**	**0.87876155**	**0.88455**	**0.881655789**	**1**
19	3	0.71127329	0.72605	0.718659835	0
19	4	0.61395046	0.62682	0.620386925	0
19	5	0.6020141	0.61167	0.606843597	0
19	6	0.58047043	0.58701	0.583741436	0
19	7	0.60487009	0.6108	0.607833277	0
19	8	0.58441557	0.58791	0.586161686	0
19	9	0.58229263	0.58469	0.58349144	0
20	**2**	**0.80364169**	**0.81437**	**0.809005776**	**1**
20	3	0.74081231	0.75207	0.746442605	0
20	4	0.67047411	0.67918	0.674826276	0
20	5	0.62816736	0.6372	0.632682782	0
20	6	0.63267728	0.6355	0.634087084	0
20	7	0.62252314	0.62147	0.621997165	0
20	8	0.63599002	0.6353	0.635646793	0
20	9	0.6303165	0.62954	0.62992993	0

s.no	image	Trainable Weka Segmentation	K=8	K auto assign
1				
2				
3				
4				

5				
6				
7				
8				
9				
10				
11				
12				
13				

14				
15				
16				
17				
18				
19				
20				

The following images are obtained from running the algorithm on very few sample of banana leafimages.

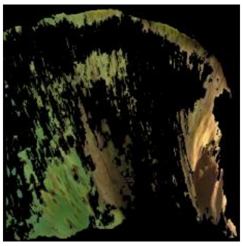


Fig1.

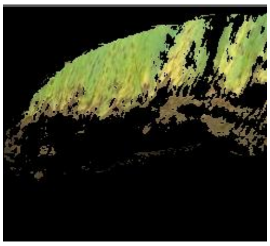


Fig 2

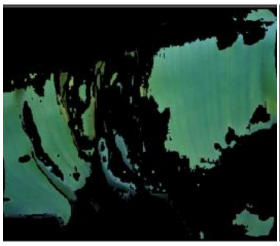


Fig 3

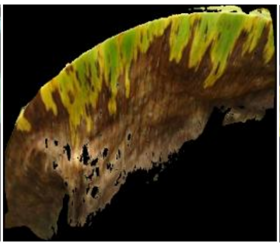


Fig 4

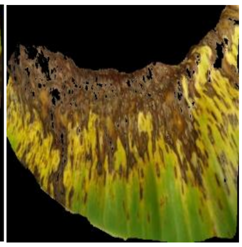


Fig 5

From the figures Fig 1,2,3,4 it can be observed that the segmentation has some errors varying upon the brightness and contrast values of the colors. The future work would be on improvising the Dev algorithm using HSV values for better segmentation.

VI. CONCLUSION

The novel algorithm proposed in this paper can be used in image data mining areas and in whichever area where segmentation and clustering is required. The major difficulty in k-means automatic selection of k has been overcome in this paper. This algorithm maybe applied in convolutional neural network as one of the hidden layer for image classification tasks.

REFERENCES

1. Mohammad Hadi Yousofi, Mahdi Esmaeili, Majide Sadat Sharifian. A Study on Image Mining; Its *Importance and Challenges. *American Journal of Software Engineering and Applications.* Special Issue: Academic Research for Multidisciplinary. Vol. 5, No. 3-1, 2016, pp. 5-9. https://doi.org//10.11648/j.ajsea.s.2016050301.12
2. Ng, H.P., Ong, S.H., Foong, K., Goh, P.M., & Nowinski, W.L. (2006). Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm. 2006 IEEE Southwest Symposium on Image Analysis and Interpretation, 61-65. https://doi.org//10.1109/ssiai.2006.1633722
3. Pallavi Purohit and Ritesh Joshi(2013), A New Efficient Approach towards k-means Clustering Algorithm, In International Journal of Computer Applications , (0975-8887), vol. 65, no. 11, March (2013).
4. Alan Jose, S. Ravi and M. Sambath (2014) , Brain Tumor Segmentation using K-means Clustering and Fuzzy C -means Algorithm and its Area Calculation. In International Journal of Innovative Research in Computer and Communication Engineering , vol. 2, issue 2, March (2014).
5. Madhu Yedla, Srinivasa Rao Pathakota and T. M. Srinivasa, Enhanced K -means Clustering Algorithm with Improved Initial Center, In International Journal of Science and Information Technologies , vol. 1(2), pp. 121–125, (2010)
6. Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. Procedia Computer Science, 54, 764–771. https://doi.org//10.1016/j.procs.2015.06.090
7. Dhaygude Sanjay B, Kumbhar Nitin P. Agricultural plant leaf disease detection using image processing. Int J Adv Res Electr Electron Instrum Eng 2013;2(1).
8. Mrunalini R Badnakhe, Deshmukh Prashant R. An application of K-means clustering and artificial intelligence in pattern recognition for crop diseases. Int Conf Adv Inf Technol 2011;20. 2011 IPCSIT.
9. Bashir Sabah, Sharma Navdeep. Remote area plant disease detection using image processing. IOSR J Electron Commun Eng 2012;2(6):31–4. ISSN: 2278-2834.
10. V. Singh, A.K. Misra, Detection of Plant Leaf Diseases Using Image Segmentation and Soft Computing Techniques, Information Processing in Agriculture(2016), doi: http://dx.doi.org/10.1016/j.inpa.2016.10.005
11. https://github.com/godliver/source-code-BBW-BBS/tree/master/bbs
12. http://stanford.edu/~cpiech/cs221/handouts/kmeans.html .
13. https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set
14. http://imagej.net/Trainable_Weka_Segmentation
15. Ignacio Arganda-Carreras, Verena Kaynig, Curtis Rueden, Kevin W Eliceiri, Johannes Schindelin, Albert Cardona, H Sebastian Seung, Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification, Bioinformatics, Volume 33, Issue 15, 01 August 2017,Pages-2424–2426, https://doi.org/10.1093/bioinformatics/btx180

AUTHORS PROFILE

Srivalli Devi S., is a research scholar, currently doing her Ph.D at the PG & Research Dept. of Computer Science, at Chikkanna Govt. Arts College, Tirupur, Tamilnadu,India. Her areas of interest include data mining, machine learning, cryptography and robotics. Her research is focused on solving agricultural problems with the help of data science.

Dr. A. Geetha, is currently heading the PG & Research Dept. of Computer Science, at Chikkanna Govt. Arts College, Tirupur -2. She has got more than 22 years of Collegiate Teaching experience. A highly motivated teacher, with the thirst for learning new concepts. She regularly updates herself by attending workshops and hands on training sessions, to keep herself abreast with latest tools and techniques. Her research area includes Data Mining and Big Data Analytics. She has got more than 50 publications in National and International Conferences and in reputed journals.