Autonomous Wheat Seed Type Classifier System

Ahmad Reza Parnian
MSc Student
Computer Engineering and IT Department
Shiraz University of Technology

Reza Javidan
Assistant Professor
Computer Engineering and IT Department
Shiraz University of Technology

ABSTRACT
After harvesting wheat, the main concern is classifying wheat seeds according to their quality, size, variety and etc. there are different procedures to measure parameters and analyzing wheat seeds but they are time-consuming and error-prone. An automated system is developed being capable to analyze and classify wheat seeds faster with higher confidence level based on defined attributes, the system uses popular K-means clustering algorithm. The base of K-means is established on squared error. Several points are given as inputs to algorithm and then they are assigned to k clusters according to distance to the centroids, each point is included in cluster which centroid is nearest to that point. A wheat dataset taken from UCI Machine Learning Repository is considered by k-means algorithm and results are analyzed. The experimental results on prototype data show the effectiveness of the proposed method.

General Terms
Autonomous Systems, Agricultural Technologies.

Keywords
Smart systems; clustering; K-means; wheat seed; UCI repository

1. INTRODUCTION
Smart systems are known as devices incorporating operation of sensing, actuation and control. Smart systems are able to analyze a situation and make decision based on available data leading to performing appropriate task. Sensors, command and control units, information transmitter and actuators are main components of smart systems. Smart systems have been used in different areas such as agriculture. Categorizing wheat seeds based on quality and other metrics is an intricate operation that has been automated by smart systems. Several systems developed in this case have troubles such as performing the task in error or expensive with respect to energy consuming, they use imaging techniques like scanning microscopy or laser technology which lead to destruction and considerable cost. Systems using imaging techniques are handled with image classification.

Clustering is assigning objects to some groups; an object is assigned to a group in which the most object similarity exists [1]. As any object is defined by some attributes, attributes differences can be criteria to classify objects, each attribute is counted as a dimension so an object is multi-dimensional attribute vector, and the goal is to place such an object in to a group in which it has the most similarity to other objects in terms of attribute values. When there is a dataset of objects intended to be clustered in determined count of groups, several algorithm may suit and the one which is most popular, familiar and widely used is K-means [2]. MATLAB is a programing environment supporting K-means algorithm. In this work a system using K-means clustering algorithm is proposed to categorize wheat seeds. An approach is provided to categorize harvested wheat seeds into groups according to quality, type, size or any desirable characteristic. K-means algorithm detects variability in attribute values and put the seeds with inconsiderable difference in attributes value in same cluster and eventually seeds are classified to cluster with the most inter-cluster similarity.

The organization of the paper is as follows: Section 2 presents a brief description of related works, K-means clustering is presented in Section 3, Section 4 is about the proposed method, experimental results are brought in Section 5 and Section 6 contains conclusion.

2. RELATED WORKS
L. Lin and et al [3] introduced a method based on fuzzy theory by considering the characteristics of wheat seed which helps in recognition the seed type. They used Tabu search in Fuzzy C-Means Clustering in phase of learning. The proposed method converges without any concern about the locality and the sensitivity of Fuzzy C-means clustering initial condition.

In [4] a system required for seed quality detection is developed using counting algorithms of particle images. Image processing techniques is used to determine the grain quality. The system is capable of extracting feature parameters and data analysis with a low cost and a high efficiency.

M. R. Neuman and et al [5] developed a workstation assisting in cereal grain inspection for classifying purposes video colorimetry methodology is proposed to help measuring color of cereal grains. Mean, variance and kurtosis are the characteristics measured for each kernel.

3. K-MEANS CLUSTRING
K-means is an unsupervised algorithm distributing data in to k clusters. Each cluster uses a concept named centroid; each point in dataset is classified into a cluster whose centroid has minimum distance to it [6]. The algorithm consists of two main phases, in the first phase batch updates happen since in each iteration after recalculation of cluster centroids, points are reassigned to their nearest cluster centroid all at once, convergence to solution may not occur in this phase meaning that reaching a local minimum is possible and is more likely for small datasets, first phase is fast and estimates a solution being a starting point for the next phase. The second phase uses online updates, in this phase points are individually reassigned leading to reduction in the sum of distances. After each reassignment cluster centroids are recomputed. In the second phase each iteration includes one pass through all the points, here convergence to a local minimum is possible although local minima with lower total sum of distances may exists. In this case global minimum can be found by running algorithm several times with random starting points [7], [8].

At the start of the algorithm k initial centroids are taken, they can be selected from dataset or k randomly points can be used as centroids [9], [7]. The next operation is calculating each

point distance to centroids and assigning the point to a cluster with nearest centroid to point. Now there are k clusters with k centroids to which points are assigned. Centroids must be updated, so in all clusters a mean of each attribute value is calculated and is assigned to centroid respective attribute [10]. All attributes value of each cluster centroid is updated by mean of respective attribute value of all points in same cluster. After that for each point, distance to each centroid is calculated and assigning the point to cluster with the minimum centroid distance is done, the process of updating centroids and reassigning points iterates until no point is assigned to a new cluster. Following steps explain k-mean clustering algorithm in brief.

1. Determining k (number of clusters)
2. Initializing k centroid
3. Calculating distances to centroid and assigning to a cluster according to distance of point to cluster centroid.
4. Updating centroid attributes value
5. Repeating steps 3 and 4 till no point is reassigned

Dataset consists of points, point i is a vector (a_{i1}, a_{i2}, a_{i3}… a_{in}) where a_{in} represents the nth attribute value of ith point. At the beginning of k-means algorithm k points in dataset are taken as centroids or k vectors are created as centroids by setting their attributes random values. K-means uses different distance metrics to perform step 3. The most widely used distance metric is the Euclidean distance. Assuming having p_i (p_{i1}, p_{i2}, p_{i3} … p_{in}) and p_j (p_{j1}, p_{j2}, p_{j3} … p_{jn}) the Euclidean distance is calculated as below:

$$D_{\text{Euclidean}}(p_i, p_j) = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2} \quad (1)$$

Euclidean distance metric is used to define a nearest centroid to a point and the point is assigned to a cluster with the nearest centroid [11]. Another distance measure is city-block which is calculated by sums of attributes absolute differences. City-block is called Manhattan metric too. The following equation calculates city-block distance between p_i and p_j.

$$D_{\text{city-block}}(p_i, p_j) = |p_{i1} - p_{j1}| + |p_{i2} - p_{j2}| + \ldots + |p_{in} - p_{jn}| \quad (2)$$

Cosine and correlation are the other distance measure metrics [12]. After each point is assigned to a cluster, new centroids can be defined, assuming vector C_i (c_{i1}, c_{i2}, c_{i3} … c_{in}) representing a centroid and i is corresponding cluster which contains n points, c_{im} will be updated as below:

$$C_{im} = \frac{\sum_{k=1}^{n} p_{km}}{n} \quad (3)$$

Where p is the points in cluster, this is what is done in step 4. Iterating steps 3 and 4 finishes when no data point changes cluster [13].

4. THE PROPOSED METHOD

Wheat seed clustering system is based on K-means clustering algorithm and the default Euclidean distance metric is used. The dataset is given to MATLAB. MATLAB contains a lot of toolboxes used in different applications. There are toolboxes used in Bioinformatics, image processing, Neural Network, Data Acquisition, financial cases and etc. Statistics toolbox contains a lot of functions which are useful in clustering [8]. kmeans function is used from statistics toolbox which is given two arguments. The first one is the dataset prepared for clustering and the second is the number of the clusters the data is going to be classified to. Dataset is given to the function as a matrix, the rows are data points and columns are attributes so a dataset with n points and n attribute becomes a p-by-n matrix argument. Function kmeans returns some vectors and matrix as outputs. A p-by-1 vector returned by kmeans shows into which cluster, points are classified and a k-by-p matrix contains the value of each centroid attribute. kmeans returns within-cluster sums of point-to-centroid distances and distances from each point to every centroid too.

UCI Machine Learning Repository provides a large collection of database and seeds-dataset has been taken from that repository. This dataset comprises randomly selected kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian. There are 70 elements of each variety. A non-destructive soft X-ray technique is used to detect internal kernel structure with a high quality while other sophisticated imaging techniques like scanning microscopy or laser technology are considerably more expensive. Seeds-dataset is Multivariate, consisting 210 instances. Seven geometric parameters of wheat kernel are used as real-valued attributes organizing an instance. These 7 attributes are:

1. area A,
2. perimeter P
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient,
7. length of kernel groove.

Dataset contains 210 records of 3 kinds of wheat seeds specification; there are 70 points of each kind. All the records are given to K-means. The algorithm is capable of clustering data points correctly. K-means clustering results may often depend on starting points and reaching to local minimum is possible when reassigning each data point to new cluster causes an increase in total sum of point to centroid distances although a better clustering solution exists [7]. Function kmeans can solve this problem by getting another argument called replicates; it is an integer number specifies how many times algorithm should be run with a new starting point [8].

5. EXPERIMENTAL RESULTS

In this part, results of testing system with valid seed-dataset received from UCI Machine Learning Repository are presented. Dataset is given as a 210*7 matrix and number 3 is taken as the cluster count. Function kmeans receives at least dataset and number of cluster as arguments. There are some other optional arguments. As there are three clusters, three centroids is needed so three records are randomly selected as cluster centroids. After running the algorithm the following is gotten as a part of result:

iter	phase	num	sum
1	1	210	601.059
2	1	8	587.985
3	1	2	587.319
4	2	0	587.319

4 iterations, total sum of distances = 587.319

'iter' shows that how many times algorithm is iterated to classify points and 'phase' is the number of phase. 'num' represents number of points assigned to new clusters and

'sum' is total summation of distances between points and their cluster centroids. At the start of operation no point is clustered so first iteration consists of 210 assignments. In second iteration there are 8 points reassigned considering that sum of distances decreases. Two reassigning happen in the third iteration. In this case the solution is found in the first phase so no reassignment is done in second phase. Just as illustrated in figure 1 number of reassigned points decrease after each iteration until reaches zero implying that minimum sum of distances is earned.

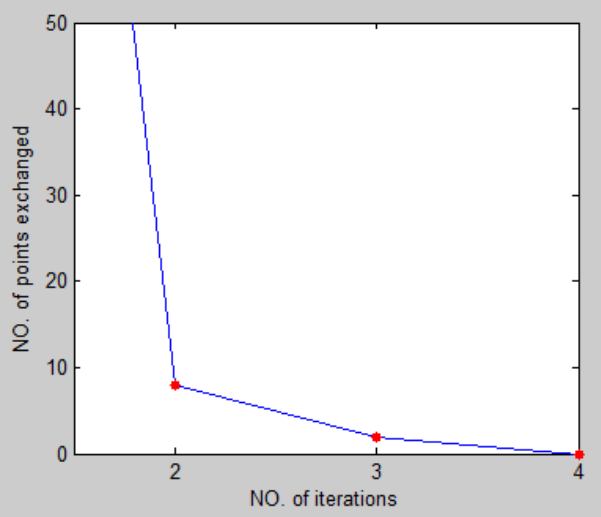

Figure 1. Number of points assigned in iterations

Sum of distances has the maximum value at the start of algorithm. During iterations sum of distances decreases since points are assigned to the appropriate clusters [14]. The minimum possible value is found when algorithm finishes. Figure 2 illustrates the changes of total sum of distances during algorithm run.

Seeds-dataset points contain seven attributes and clustering is done base on all the seven attributes so a seven-dimensional diagram illustrates the clustered point the best but if assuming the dataset with just the first three attributes (area, perimeter and compactness) clustering can be exhibited as Figure 3. Each color represents a cluster and each colored dot is a point.

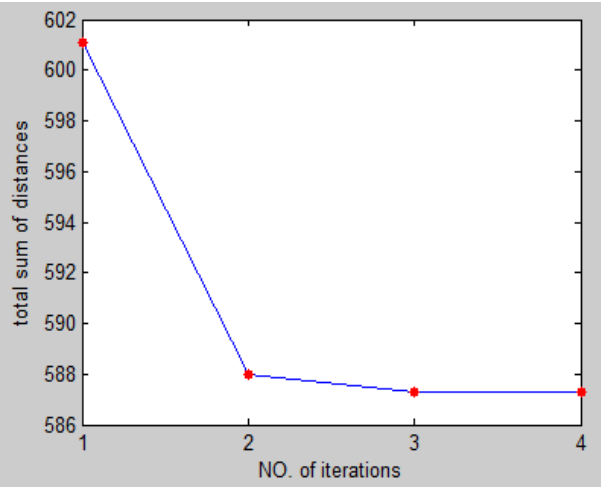

Figure 2. Total sum of distance versus iteration

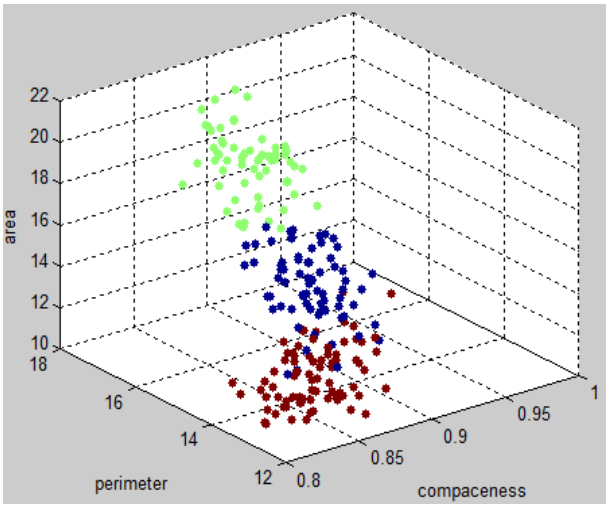

Figure 3. Points in 3D space

6. CONCLUSIONS

A system for clustering wheat seeds is proposed, the results of the experimented system with wheat dataset taken from UCI machine learning repository show a high level of accuracy and success. The system is capable of clustering approximately all the seeds correctly. The profiting K-means algorithm leads to fast and efficient operation. In addition, it provides a more economical solution.

7. REFERENCES

[1] J. Wu, Advancs in K-means Clustering: A Data Mining Thinking, Springer, 2012.

[2] K. Alsabti, S. Ranka and V. Singh, "An Efficient k-means Clustering Algorithm", Proc, First Workshop High Performance Data Mining, Mar 1998.

[3] L.Lin and L.Suhua, "Wheat Cultivar Classifications Based on Tabu Search and Fuzzy C-means Clustering Algorithm", Fourth International Conference on Computational and Information Sciences, pp. 493-496, Aug 2012.

[4] D. Liying, Z. Genshan, L. Xuning, S. Wei, L. Aiqin and C. Weihua, "Design and Realization of Grain Seed Quality Testing System Based on Particle Image Processing Technology", International Conference on Computer Science and Electronics Engineering, vol. 3, pp. 61-65, March 2012.

[5] M.R. Neuman, E. Shwedyk and W. Bushuk, "A PC-based colour image processing system for wheat grain grading", International Conference on Image Processing and its Applications, pp. 242-246, Jul 1989.

[6] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[7] P. S. Bradley and U. Fayyad, "Refining Initial Points for K-means Clustering", Proc. 15th Int'l Conf. Machine Learning, pp. 91-99, 1998.

[8] http://www.mathworks.com,"StatisticsToolbox:K-means Clustering", r2013a.

[9] R. C. Dubes and A. K. Jain, Algorithms for Clustering Data, Prentice Hall, 1988.

[10] K. Wagsta, C. Cardie, S. Rogers and S. Schroedl, "Constrained K-means Clustering with Background Knowledge", Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577-584, 2001.

[11] S. Arora, P. Raghavan, and S. Rao, "Approximation Schemes for Euclidean k-median and Related Problems", Proc. 30th Ann. ACM Symp. Theory of Computing, pp. 106-113, May 1998.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "The Analysis of a Simple k-means Clustering Algorithm", sixteenth annual symposium on Computational geometry, pp. 100-109, Jan 2000.

[13] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, " An Efficient k-Means Clustering Algorithm:Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, NO. 7, pp. 881-892, July 2002.

[14] V. Faber, "Clustering and the Continuous k-means Algorithm", Los Alamos Science, vol. 22, pp. 138-144, 1994.