SUGGEST CLUSTERING ALGORITHM FOR FEATURES ANALYSIS OF X-RAY IMAGES

Dr. Samaher Hussein Ali[*1], **Hawraa Shareef Hamza** [*2] **Dr. Ghaidaa Al-Sultany** [*3]

samaher@itnet.uobabylon.edu.iq hawraa_shareef@yahoo.com ghaidaa.almulla@gmail.com

[1,2,3] *Software Department, Information Technology College,Babylon University, Iraq*

ABSTRACT

This paper presents a new clustering algorithm based on similarity and dissimilarity measurements which applied to X-Ray Images. The main phases of that paper are (pre-processing phase which including feature extraction and find correlation among features, processing phase include applying the new algorithm called similarity and dissimilarity clustering algorithm “SDCA” to find the best number of clusters, and the final phase called post processing phase that including generation ANDing Rules and verification of these rules). In this paper, we are generating knowledge base as rules satisfy the correct rate equal 100 %. In addition, this work compares among four types of clustering algorithm include (Fuzzy C-means clustering algorithm, Genetic algorithm, Swarm Particle Algorithm and SDCA). Add to that, it proves by results SDCA is better from (Number of epochs, Final number of clusters, Fitness Function and Accuracy) than other three clustering methods used in compare.

Keyword: Clustering, Similarity and Dissimilarity measures, GA, FCM, POS and SDCA.

1. INTRODUCTION

Clustering is the process of grouping the data into classes or *clusters*, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.

Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. In this paper, we study the requirements of clustering methods for large amounts of data. We explain how to compute dissimilarities between objects represented by various attribute or variable types.

While, the cluster analysis can be define as finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. Figure 1 show the main idea of clustering.

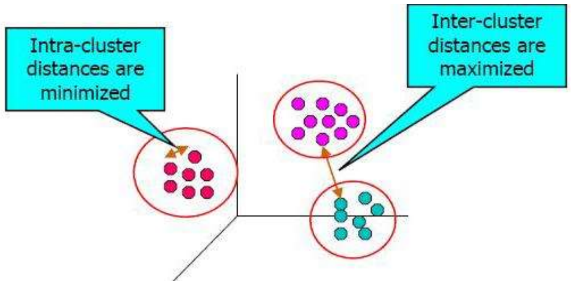
Intra-cluster distances are minimized
Inter-cluster distances are maximized

Fig.1. the Clustering Principle

As a result, Clustering [1][2] is a popular unsupervised pattern classification technique which partitions the input space into K regions based on some similarity or dissimilarity metric. The number of partitions or clusters may or may not be known a priori. Let the input space S be represented by n points {x1, x2,…,xn}, and the K clusters be represented by C1 ,C2 ,…. , CK . Then

(i) $C_i \neq 0$ for i=1, 2,… , K

(ii) $C_i \cap C_j = 0$ for i=1,2,….., K and j=1,2,….., K and i≠j

(iii) $\overset{K}{\underset{i=1}{\mathrm{U}}} C_i = S$

In general the clustering methods include the following elements [3]:

(i) Pattern representation: determines number of clusters, number of variable vectors and number of features in the feature vector.

(ii) Feature selection: defines a subset of features to use in clustering process.

(iii) Data abstraction: represents a process to find simple representation of clustering sets.

(iv) Assignment measure: explains how we can combine feature vectors by feature selection of one of the variable cluster. There are two types of these measures:

- Distance measures such as Euclidean distance, Minkowski distance.
- Similarity measures such as Vector inner product.

While, the main Types of Clusters:

(i) Well-separated clusters: A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

(ii) Center-based clusters: A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster**.** The center of a cluster is often a **centroid,** the average of all the points in the cluster, or a **medoid**, the most "representative" point of a cluster

(iii) Contiguous clusters (Nearest neighbor or Transitive): A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

(iv) Density-based clusters: A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

(v) Property or Conceptual: Finds clusters that share some common property or represent a particular concept.

(vi) Described by an Objective Function:

The remind of this paper is organization as follow, Section 2 show overview of the previous works and compare among it. Section 3 presents the main tools used. Section 4 show the suggest clustering algorithm. Section 5 shows the experiment result and compare the result of suggest algorithm with GA, FCM and POS. Finally, we discuss the results and future work in the conclusion section.

2. OVERVIEW

The clustering techniques are considering one of the popular methodologies in many fields. Therefore, in this section we show a set of previous works that using a clustering in different aspects and compare among it.

kuang- et al.,2005 [4] compares four similarity measures such as the city block (L1-norm), Euclidean (L2-norm), normalized correlation coefficient, and simplified grey relational grade for clustering of QRS complexes. Performances of the measures include classification accuracy, threshold value selection, noise robustness, execution time, and the capability of automated selection of templates. The clustering algorithm used is the so-called two-step unsupervised method. The best out of the 10 independent runs of the clustering algorithm with randomly selected initial template beat for each run is used to compare the performances of each similarity measure.

Silvia, 2007 [5], present approach yields interesting hints for potentially relevant combinations of epidemiologic and genetic risk or beneficial factors for sporadic breast cancer. Furthermore, it provides a general insight into relationships between the considered variables that may also be useful for the generation of biological hypotheses. A number of epidemiologic variables can be detected that do not contribute to differences between cases and controls and might be omitted in a further step of the analysis of the data.

Anna., 2008[6] compare and analyze the effectiveness of these measures in partitional clustering for text document datasets. The experiments utilize the standard K-means algorithm and report results on seven text document datasets and five distance/similarity measures that have been most commonly used in text clustering. This investigation found that except for the Euclidean distance measure, the other measures have comparable effectiveness for the partitional text document clustering task. Pearson correlation coefficient and the averaged KLD divergence measures are slightly better in that their resulting clustering.

Guadalupe, et al.,2008 [7] introduces a measure of similarity between two clustering of the same dataset produced by two different algorithms, or even the same algorithm K-means. Experimental results pertaining to the text categorization problem of a Portuguese corpus are presented, as well as results on the well-known benchmark IRIS dataset. The results show that Repeated Bisection and Direct hierarchical clustering algorithms consistently produced clusters that are most similar to expert labeled categories for both smaller data sets with fewer features (Iris) and large dataset (translated Portuguese-English corpus).

Mrutyunjaya et al.,2008 [8] propose some clustering algorithms such as K-Means and Fuzzy c-Means for network intrusion detection. And built a system which created clusters from its input data, then automatically labeled clusters as containing either normal or anomalous data instances, and finally used these clusters to classify network data instances as either normal or anomalous. Both the training and testing was done using KDDCup'99 data, which is a very popular and widely used intrusion attack dataset. This paper show that FCM works very efficiently in obtaining compact well separated n-clusters to detect network intrusions.

Thomas et al.,2010 [9] explore the implications of correlations for the actual and estimated precision of least squares estimators. and show that with equal sized clusters, if the covariate of interest is randomly assigned at the cluster level, only accounting for non-zero covariances at the cluster level, and ignoring correlations between clusters, leads to valid standard errors and confidence intervals

K.P.Satya et al.,2012 [10], we introduce to develop a novel hierarchal algorithm for document clustering which provides maximum efficiency and performance.. Experiments in both public data and document clustering data show that this approach can improve the efficiency of clustering and save computing time.

Saurabh et al.,2012 [11] introduce a novel multi-viewpoint based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved.

As a result, we can compare among these works base on the following points (dataset used, tools used in clustering, similarity or dissimilarity measures) as explain in Table 1.

Table 1: Comparison of the previous works

Researcher	Data set	Tools	Similarity
kuang-chiung , et al., 2005[4]	QRS complexes	two-step unsupervised method	city block (L1-norm), Euclidean (L2-norm), normalized correlation coefficient
Silvia , et al., 2007[5]	Single Nucleotide Polymorphisms	cluster analysis	Flexible Matching Coefficient, Pearson's Corrected Coefficient of Contingency, mixed similarity coefficient
Anna Huang 2008 [6]	Text documents	K-means algorithm	cosine similarity and the Jaccard correlation coefficient
Guadalupe , et al., 2008 [7]	Text documents, Iris data set.	K-means	Euclidean (centroids) distances and Pearson correlation coefficient
Mrutyunjaya , et al., 2008 [8]	Network Data set	k-means, fuzzy c-means, mf plot, Roc	Euclidean Distances
Thomas , et al., 2010[9]	illustrative purposes census data	randomization inference	spatial correlations
K.P.Satya et al., 2012[10]	Text document	hierarchal algorithm	Cosine similarity
Saurab et al.,	Text document	Hierarchical Clustering Method,	Multi-viewpoint Based Similarity Measure

3. TOOLS

3.1. CORRELATION MEASURES

We will use the correlation coefficients as an indicator to know the degree of correlation among the features. Correlation coefficient (*r*) is a statistical concept, which helps in establishing a relation between the predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values. In other words, it summarizes the strength of relationship between two vectors. The mathematical formula for computing *r* is [12]:

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{N\left(\sum x^2\right) - (\sum x)^2}\sqrt{N\left(\sum y^2\right) - (\sum y)^2}}, \qquad \text{.....(1)}$$

Where *N* represent the vector size, the value of *r* is such that -1 <= *r* <=+1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively [12].

- Positive correlation: If *x* and *y* have a strong positive linear correlation, *r* is close to +1. An *r* value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between *x* and *y* vectors such that as values for *x* increase, values for *y* also increase.
- Negative correlation: If *x* and *y* have a strong negative linear correlation, *r* is close to -1. An *r* value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between *x* and *y* such that as values for *x* increase, values for *y* decrease.
- No correlation: If there is no linear correlation or a weak linear correlation, *r* is close to 0. A value near zero means that there is a random, nonlinear relationship between the two vectors.
- A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If *r* = +1, the slope of this line is positive. If *r* = -1, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. These values can vary based upon the "type" of data being examined.

Several different correlation coefficients can be calculated. The two most commonly used are Pearson's correlation coefficient and patial correlation coefficient [13]. The **partial correlation** is [14]

$$\hat{p}_{XY.Z} = \frac{N\sum_{i=1}^{N} r_{X,i} r_{Y,i} - \sum_{i=1}^{N} r_{X,i} \sum_{i=1}^{N} r_{Y,i}}{\sqrt{N\sum_{i=1}^{N} r^2{}_{X,i} - (\sum_{i=1}^{N} r_{X,i})^2}\sqrt{N\sum_{i=1}^{N} r^2{}_{Y,i} - (\sum_{i=1}^{N} r_{Y,i})^2}} \qquad \text{.....(2)}$$

***Pearson* correlation coefficient** is [15]

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \qquad \text{...... (3)}$$

3.2. FUZZY C-MEANS CLUSTERING ALGORITHMS

In this section, an attempt has been made to use fuzzy c-means clustering algorithms for automatically clustering a data set. This includes determination the best cluster as well as appropriate clustering of the data.

```
Procedure Fuzzy C-Means of clustering
Input: X-Ray Images.
Output: Best clustered of X-Ray images.
Preparations:
    * Fix c, $2 \leq c < d$
    * Choose any inner product norm metric for Rn.
    * Choose the termination tolerance $\delta > 0$,
        e.g between 0.01 and 0.001.
    * Fix w, $1 \leq w < \infty$, e.g. 2.
    * Initialise U (0) Є Mfc, (e.g. randomly).
Repeat
Step 1: Compute cluster prototypes:
    $\mathbf{c}_l^{(i)} = \frac{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^w \mathbf{m}_j}{\sum_{j=1}^{d} \left(u_{ij}^{(l-1)}\right)^w}, \quad 1 \leq i \leq c$
Step 2: Compute distances:
        For all clusters $1 \leq i \leq c$,
        For all data objects $1 \leq j \leq d$,
    $d_{\mathbf{A}}^2 \left(\mathbf{m}_j, \mathbf{c}_l^{(i)}\right) = \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j\right)^T \mathbf{A} \left(\mathbf{c}_l^{(i)} - \mathbf{m}_j\right)$
Step 3: Update the partition matrix:
    If $d_{\mathbf{A}} \left(\mathbf{m}_j, \mathbf{c}_l^{(i)}\right) > 0$ for $1 \leq i \leq c, 1 \leq j \leq d$,
        $u_{ij}^{(l)} = \frac{1}{\sum_{k=1}^{c} \left(d_{\mathbf{A}}^2(\mathbf{m}_j, \mathbf{c}^{(i)}) / d_{\mathbf{A}}^2(\mathbf{m}_j, \mathbf{c}^{(k)})\right)^{1/(w-1)}}$
    otherwise
    $u_{ij}^{(l)} = 0$ if $d_{\mathbf{A}} \left(\mathbf{m}_j, \mathbf{c}_l^{(i)}\right) > 0$, and $u_{ij}^{(l)} \in [0, 1]$ with $\sum_{i=1}^{c} u_{ij}^{(l)} = 1$
Until $\left\| \mathbf{U}^{(l)} - \mathbf{U}^{(l-1)} \right\| < \delta$
```

3.3 GENETIC ALGORITHM

The genetic algorithm (GA) is a heuristic used to find approximate solutions for difficult to solve problems through application of the principles of evolutionary biology to computer science. Genetic algorithms use biologically-derived techniques such as inheritance, mutation, natural selection, and recombination (or crossover). Genetic algorithms are a particular class of evolutionary algorithms. GA based clustering technique can automatically evolve the appropriate clusters number of a data set.

```
Procedure GA of clustering
 Input: X-Ray Images.
Output: Best clustered of X-Ray images.
Step 1: Set Kmin , , Kmax to min & max clusters
number expected respectively, Set MaxGen to max
iteration allowed
 Gen ← 1
 Step 2 : Population initialization:
   For each chromosome in the population
      *Generate a number K, in the range Kmin to Kmax
      *Choose Ki points(rows) randomly from the dataset
      *Distribute these points randomly in the chromosome
      *Set unfilled positions to null
 Step 3: Fitness calculation:
   For each chromosome in the population
              *Extract the Ki centers stored in it
             *Perform clustering by assigning each point to the
            cluster corresponding to the closest center
 Step 4: Elitism
 Step 5: If Gen> Maxgen GOTO Step 7.
 Step 6: Genetic Operations:
         Selection
         Single point crossover with prob. Mc
        Mutation performed with prob. $\mu m$:
            Randomly choose one position of chromosome.
            If this position is null randomly, choose a point from
               data and make it as a center
           Else make this position null
 Step 7: End.
```

3.4. SWARM PARTICLE ALGORITHM

Particle swarm optimizers (PSO) are optimization population algorithms formed after the imitation of social behavior of bird flocks [16]. PSO is generally considered to be an evolutionary computation (EC) example. genetic algorithms (GA), Genetic Programming (GP), Evolutionary Strategies (ES), and Evolutionary Programming (EP) are other examples of Evolutionary computation (EC) [17], which simulate a biological evolution approach. In a PSO system, a swarm of individuals (called particles) fly through the search space. Each particle represents a candidate solution to the optimization problem. The best position visited by particle (i.e. its own experience) and the position of the best particle in its neighborhood (i.e. the experience of neighboring particles) are affected to the position of a particle.

The best position in the neighborhood is referred to as the global best particle when the neighborhood of a particle is the entire swarm, , and the resulting algorithm is referred to as a g_{best} PSO. When smaller neighborhoods are used, the algorithm is generally referred to as a l_{best} PSO [18]. The performance of each particle is measured (i.e. how close the particle is from the global optimum) using a fitness function that changing depending on the optimization problem.

The PSO method is one of optimization methods developed for searching global optima of a nonlinear function [19]. It is inspired by the social behavior of birds and fish. The method uses group of problem solutions. Each solution consists of set of parameters and represents a point in multidimensional space. Each potential solution is called particle and the group of particles (population) is called swarm. Particles move through the search domain with a specified velocity in search of optimal solution. Each particle maintains a memory which helps it in keeping the track of its previous best position. The positions of the particles are distinguished as personal best and global best. PSO has been applied to solve a variety of optimization problems and its performance is compared with other popular stochastic search techniques like Genetic algorithms, Differential Evolution, Simulated Annealing etc. [20].

The working of the PSO may be described as: For a D_ dimensional search space the position of the ith particle is represented as $Xi = (x_{i1}, x_{i2}, \dots x_{iD})$. Each particle maintains a memory of its previous best position $P_{best} = (p_{i1}, p_{i2} \dots p_{iD})$. The best one among all the particles in the population is represented as $P_{gbest} = (p_{g1}, p_{g2} \dots p_{gD})$. The velocity of each particle is represented as $Vel_i = (vel_{i1}, vel_{i2}, \dots vel_{iD})$. In each iteration, the P vector of the particle with best fitness in the local neighborhood , designated g, and the P vector of the current particle are combined to adjust the velocity along each dimension and a new position of the particle is determined using that velocity. The two basic equations used for updating the velocity vector and position vector are given by [21] :

$$vel_{id}=w.\ vel_{id} + c_1 rand()(p_{id} - x_{id}) + c_2 Rand()(p_{gd} - x_{id}) \quad \dots\dots \quad \mathbf{(4)}$$

$$x_{id} = x_{id} + vel_{id} \quad \dots\dots\dots \quad \mathbf{(5)}$$

The first part of equation (4) represents the inertia of the previous velocity, the second part is the perception part and it tells us about the personal experience of the particle, the third part represents the cooperation among particles and is therefore named as the social component. Acceleration constants c_1, c_2 and inertia weight w are the predefined by the user and r_1, r_2 are the uniformly generated random numbers in the range of [0, 1].

The parameters of this algorithm can be stated as fellows: N_d dimension of the input, which means the number of parameters for each data vector; N_o denotes the number of data vectors to be clustered; Nc denotes the number of cluster centroids to be formed (as provided by the user),x_p denotes the x^{pth} data vector; v_j denotes the centroid vector of cluster j ; n_k is the number of data vectors in cluster j; C_j is the subset of data vectors that form cluster j.

In the context of clustering, a single particle represents the N_c cluster centroid vectors [22]. That is, each particle X_i is constructed as follows:

$$X_i = (v_{i1},...,v_{ij},...,v_{iN_c})$$

where v_{ij} refers to the j^{th} cluster centroid vector of the i^{th} particle in cluster C_{ij} . Therefore, a swarm represents a number of candidate clusters for the current data vectors. The fitness of particles is easily measured as the quantization error [45].

The fitness function= $$\frac{\sum_{j=1}^{N_C}\left[\sum_{\forall x_p \in C_{ij}} d(X_p, v_j)\right] / n_k}{K} \qquad .. ....(6)$$

where C_k is the k^{th} cluster, and N_c is the number of pixels in $\boldsymbol{C_k}$,K is the total number of cluster , d is defined in equation (7)

```
Procedure PSO of clustering
 Input: X-Ray Images.
Output: Best clustered of X-Ray images.
Step 1: Initialize each particle to contain $N_c$ randomly selected cluster centroids.
Step 2:  For $t = 1$ to $t_{max}$ do
   Step 2.1. For each particle i do
  Step 2.2. For each data vector $X_p$
       * calculate the Euclidean distance $d(X_p ,v_{ij})$ to all
         cluster centroids
       *assign $X_p$ to cluster $v_{ij}$ such that distance
         $d(X_p ,v_{ij})= \min_{\forall c=1,...,N_c} \{d(X_p - v_{ic})\}$        ...... (7)
     * calculate the fitness .
   Step 2.3. Update the global best and local best positions
   Step 2.4. Update the cluster centroids.
```

where t_{max} is the maximum number of iterations. The population-based search of the PSO algorithm reduces the effect that initial conditions which has the search starts from multiple positions in parallel opposite to the K-means algorithm. However, the K-means algorithm tends to converge faster (after less function evaluations) than the PSO, but usually with a less accurate clustering [23] , showed that the performance of the PSO clustering algorithm can further be improved by seeding the initial swarm with the result of the K-means algorithm [24]. The hybrid algorithm first executes the K-means algorithm once. In this case the K-means clustering is terminated when (1) the maximum number of iterations is exceeded, or when (2) the average change in centroids vectors is less that 0.0001 (a user specified parameter).

PSO has a set of parameters such as inertia weight, acceleration constants, swarm size etc which is to be defined by the user like all Evolutionary Algorithms. These parameters may be varied as per the complexity of the problem. Explained the basic parameters in PSO [25].

4. SIMILARITY AND DISSIMILARITY CLUSTERING ALGORITHM (SDCA)

Procedure SDCA of Clustering

Input: *X-Ray Images.*

Output: Best *clustered of X-Ray images.*

Step 1: *Extract the Features from X-Ray images.*

Step2: *Find the correlation among features to remove in active features*

Step 3: *Draw the histogram of X- Ray Images to determine the number of clusters (i.e., peak have more frequency become the seed of a cluster or center of the cluster)*

Step 4: *Consider the above resulted cluster center as the cluster center of the first Particle of the initial swarm.*

Step 5: *Generate the rest particle of the initial swarm.*

Step 6: *Initialize the individual best of each particle and the global best of the entire swarm.*

Step 7: *Evaluate the initial swam using the fitness function for each particle base on eq(6) using different similarity and dissimilarity measures (i.e., where, we find the min distance when using dissimilarity measure and max distance when using similarity measure)*

7.1 *Bray-Curtis distance (dissimilarity measure)*

$$d^{BCD}(i,j)=\frac{\sum_{k=0}^{n-1}|y_{i,k}-y_{j,k}|}{\sum_{k=0}^{n-1}(y_{i,k}-y_{j,k})}$$

7.2 *Canberra distance (dissimilarity measure)*

$$d^{CAD}(i,j)=\sum_{k=0}^{n-1}\frac{|y_{i,k}-y_{j,k}|}{|y_{i,k}|+|y_{j,k}|}$$

7.3 *Chi-Square Distance(dissimilarity measure)*

$$d(i,i)=\sqrt{\sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}}-\frac{f_{i'j}}{f_{i'.}}\right)^{2}\cdot\frac{1}{f_{.j}}}$$

7.4 *Square chord distance dissimilarity measure)*

$$D=\sum_{i=1}^{i-n}(\sqrt{x_i}-\sqrt{y_i})^2$$

7.5 Euclidean distance (dissimilarity measure)

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

7.6 Kendall distance (dissimilarity measure)

$$k(\tau_1, \tau_2) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(\tau_1, \tau_2)$$

7.7 Mahalanobis distance (dissimilarity measure)

$$D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}.$$

7.8. Cosine (similarity measure)

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\| A \| \, \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

7.9 Sokal & Michener (similarity measure)

$$\frac{2C}{N_1 + N_2} = \frac{2n_{JK}}{2n_{JK} + n_{JK} + n_{jk}}$$

7.10 Kulczynski (similarity measure)

$$\frac{C(N_1 + N_2)}{2(N_1 N_2)} = \frac{n_{JK}}{2n_J} + \frac{n_{JK}}{2n_K}$$

7.11 Dice similarity coefficient (DSC) (similarity measure)

$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

7.12 Jaccard similarity (similarity measure)

$$d^{JAS}(i, j) = \frac{j11}{j01 + j10 + j11}.$$

Step8: *Select the best individual and global best of the Swarm*

Step 9: *Update the velocity and the position of each particle using the equations (4) and (5).*

Step 10: *Obtain the optimal solution in the initial stage.*

Step 11: *Repeat step 7-step 10 until the maximum number of iterations specified.*

Step 12: *Obtain the optimal Clustering solution at the end.*

Step 13: *Generation Anding Rules*

Step 14: *Verification of the result rules.*

5. EXPERIMENTS

5.1. DESCRIPTION OF DATASETS

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is nondestructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The data set can be used for the tasks of cluster analysis. **Table 1: Description of Dataset**

Data Set Characteristics:	**Multivariate**
Attribute Characteristics:	**Real**
Associated Tasks:	**Clustering and Classification**
Number of Instances:	**210**
Number of Attributes:	**7**
Missing Values?	**N**
Area:	**Life**
Date Donated	**2012-09-29**
Number of Web Hits:	**6727**

5.2. RESULTS PRESENTS BY SDCA

Step 1: Extract the Features from X-Ray images.

Seven geometric parameters are measured:

F1: Area A,

F2: Perimeter P,

F3: Compactness C = 4*pi*A/P^2,

F4: Length of kernel,

F5: Width of kernel,

F6.:Asymmetry coefficient

F7:length of kernel groove.

All of these parameters were real-valued continuous.

Step2: Find the correlation coefficient (r) and the correlation Pearson among features to remove in active features as explain in the tables 2 and 3. Where, that tables show F6 not correct with the other features therefore, we remove that feature.

Table 2: Correlation Coefficient (*r*)

Var.	F1	F2	F3	F4	F5	F6	F7
F1	**1.000**	0.994	0.608	0.950	0.971	-0.230	0.864
F2	0.994	**1.000**	0.529	0.972	0.945	-0.217	0.891
F3	0.608	0.529	**1.000**	0.368	0.762	-0.331	0.227
F4	0.950	0.972	0.368	**1.000**	0.860	-0.172	0.933
F5	0.971	0.945	0.762	0.860	**1.000**	-0.258	0.749
F6	-0.230	-0.217	-0.331	-0.172	-0.258	**1.000**	-0.011
F7	0.864	0.891	0.227	0.933	0.749	-0.011	**1.000**

Table 3: Correlation matrix (Pearson)

Var.	F1	F2	F3	F4	F5	F6	F7
F1	1	0.994	0.608	0.950	0.971	-0.230	0.864
F2	0.994	1	0.529	0.972	0.945	-0.217	0.891
F3	0.608	0.529	1	0.368	0.762	-0.331	0.227
F4	0.950	0.972	0.368	1	0.860	-0.172	0.933
F5	0.971	0.945	0.762	0.860	1	-0.258	0.749
F6	-0.230	-0.217	-0.331	-0.172	-0.258	1	-0.011
F7	0.864	0.891	0.227	0.933	0.749	-0.011	1

Step 3: Draw the histogram of X- Ray Images to determine the number of clusters (i.e., peak has more frequency become the seed of a cluster or center of the cluster) as explain in Figure 2.

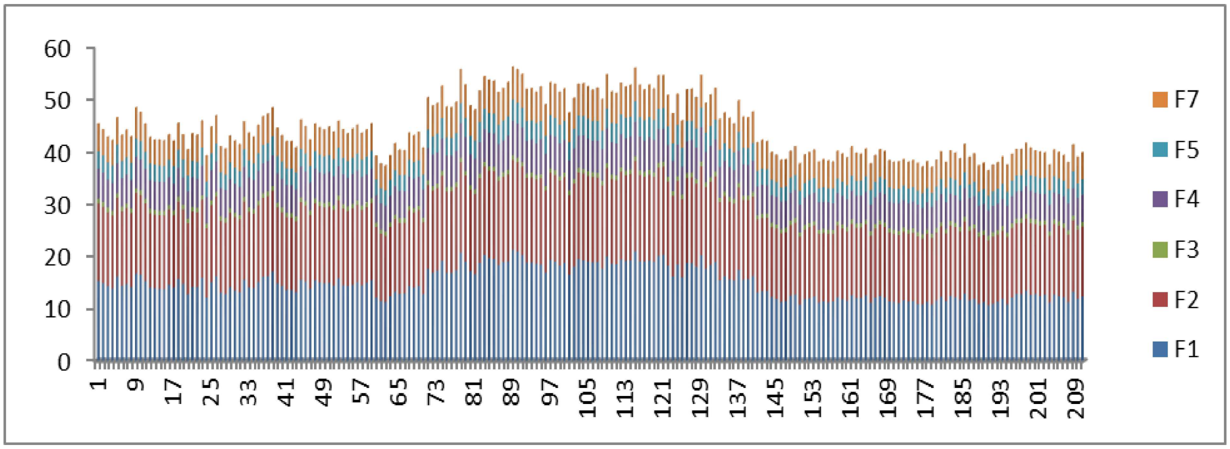

Fig.2. The Histogram of the Dataset

Step 4: Consider the above resulted cluster centre as the cluster centred of the first Particle of the initial swarm. As explain in Table 4.

Table 4: Initial Custer Centroids

Center	F1	F2	F3	F4	F5	F7
1	15.132	14.673	0.874	5.652	3.302	5.445
2	14.883	14.607	0.866	5.673	3.241	5.467
3	14.914	14.583	0.871	5.648	3.269	5.415
4	14.404	14.384	0.868	5.570	3.195	5.335
5	14.900	14.557	0.874	5.615	3.279	5.392
6	15.711	14.968	0.881	5.739	3.400	5.441
7	13.564	13.932	0.877	5.381	3.148	4.957
8	12.026	13.317	0.851	5.257	2.887	5.149
9	18.924	16.383	0.886	6.238	3.745	6.097
10	11.840	13.228	0.850	5.217	2.845	5.095

Step 5: Obtain the optimal solution by applying the stages of propose algorithm, we get the following results as show in table 5. While, table 6 represents the distributed of records base on the final centred generate after complete the max number of iterations.

Step 6: Generation Rules and Verification of the Result Rules. Figure 2 shows the surface of the final clusters base on the features more important (F1, F2, and F3).

Table 5: Central objects:

Final center	F1	F2	F3	F4	F5	F7
1	15.490	14.940	0.872	5.757	3.371	5.228
2	18.950	16.420	0.883	6.248	3.755	6.148
3	12.110	13.270	0.864	5.236	2.975	5.012

Table 6: Distributed of Records Base on the Final Centred

from \ to	1	2	3	Total
1	63	1	6	70
2	2	68	0	70
3	15	0	55	70
Total	80	69	61	210

Rule 1:If F4 in [5.722, 5.789] and F3 in [0.827, 0.87] and F5 in [2.693, 3.27 4] and F2 in [0.765, 3.779] and F7 in [4.519, 5.576] then CLASS = 1 in 100% of cases

Rule 2: If F1 in [13.84, 15.185] and F5 in [3.274, 3.683] and F2 in [0.765, 3.779] and F7 in [4.519, 5.576] then CLASS = 1 in 100% of cases

Rule 3: If F3 in [0.871, 0.893] and F1 in [15.185, 17.08] and F5 in [3.274, 3.683] and F2 in [0.765, 3.779] and F7 in [4.519, 5.576]then CLASS = 1 in 100% of cases

Rule 4: If F5 in [3.285, 3.434] and F1 in [3.779, 8.456] and F7 in [4.519, 5.576] then CLASS = 1 in 100% of cases

Rule 5: If F1 in [14.875, 15.26]and F5 in [2.63, 3.285] and F3 in [3.779, 8.456] and F7 in [4.519, 5.576] then CLASS = 1 in 100% of cases

Rule 6: If F2 in [5.085, 8.456] and F1 in [10.59, 14.875] and F5 in [2.63, 3.285] and F7 in [4.519, 5.576] then CLASS = 3 in 100% of cases

Rule 7: If F1 in [17.59, 21.18[and F7 in [5.576, 6.55[then CLASS = 2 in 100% of cases

Rule 8:If F3 in [3.914, 5.532] and F1 in [15.38, 17.59] and F7 in [5.576, 6.55] then CLASS = 2 in 100% of cases

Rule 9:If F3 in [0.853, 0.87] and F4 in [2.04, 3.914]and F1 in [15.38, 17.59]and F7 in [5.576, 6.55] then CLASS = 2 in 100% of cases

Rule 10:If F4 in [6.099, 6.145] and F3 in [0.87, 0.889] and F5 in [2.04, 3.914] and F1 in [15.38, 17.59] and F7 in [5.576, 6.55] then CLASS = 2 in 100% of cases

The following Table 7 show, the proposed methodology work best than another (i.e., FCM, GA, PSO)

Table 7: Compare among the Three Clustering Methods

Algorithm name	Number of epoch	Final number of cluster	Fitness Function	Accuracy
FCM	30	4	0.736	77%
GA	25	4	0.799	87%
PSO	14	3	0.892	95%
SDCA	8	3	0.991	100%

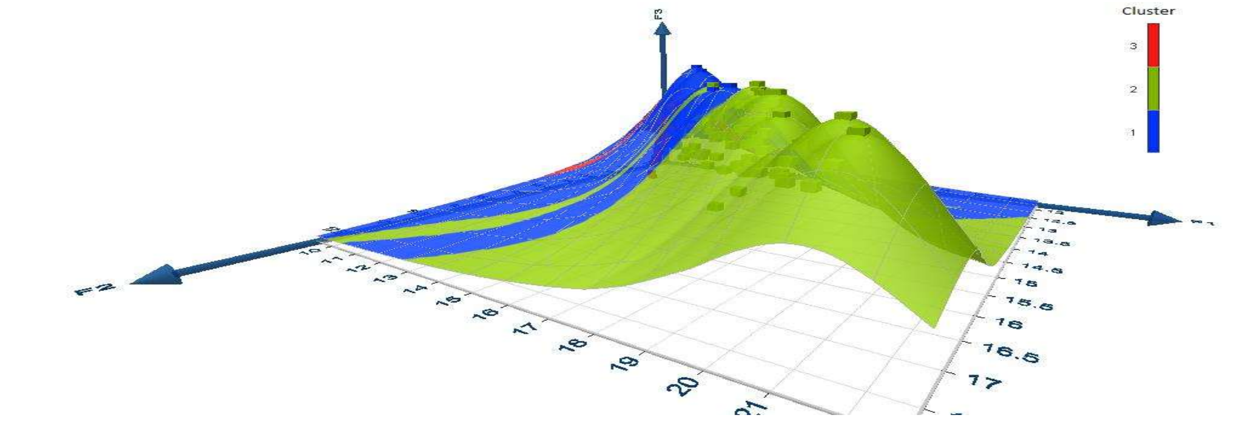

Fig.2. Surface of the final clusters base on the features more important (F1, F2, and F3)

6. CONCLUSIONS

The main idea of this paper is to present a new clustering algorithm based on similarity and dissimilarity measurements which applied to X-Ray Images. The main phase of that paper are (pre-processing phase which including Feature Extraction and find correlation among features, processing phase include applying the new algorithm called similarity and dissimilarity clustering algorithm "SDCA" to find the best number of clusters, and the final phase called post processing phase that including generation ANDing Rules and verification of these rules).

This paper show the use of the SDCA clustering technique proves faster and accrue clustering process. Also it shows that SDCA is most suitable tool to construct rules and verify them. We so that the extracted knowledge is clearly understandable and have a good generality.

This paper proves by results SDCA is better from (Number of epochs, Final number of clusters, Fitness Function and Accuracy) than other clustering methods (i.e., Fuzzy C-means clustering algorithm, Genetic algorithm, Swarm Particle Algorithm) as explained in table 7. Also, it is show when the dataset have high correlation the best measure can use with it is ***cos as similarity measure*** and ***Canberra distance as dissimilarity measure***. While, if we not verification from the correlation, then best measure can use with it is ***Dice similarity coefficient*** as similarity measure and ***Mahalanobis distance*** as dissimilarity measure. These measures conclude base on the X-Ray Images dataset. But not general for every datasets. Future work, we can design methodology to predicate the best similarity or dissimilarity measure can be used with all databases.

REFERENCES

[1] Sanghamitra.B And Ujjwal.M, "Genetic Clustering For Automatic Evolution OF Clusters And Application To Image Classification", The Journal Of The Pattern Recognition Socirty, Vol.35, PP. 1197-1208,2002.
Site:Http://Www.Elsevier.Com\Locate\Patcog

[2] Scott.E, "Computer Vision And Image Processing :A Practical Approach Using CVIP Tools", Prentice-Hall ,New Jersey ,1998.

[3] Ahmed.K, "A Genetic Clustering For Image Segmentation",M.Sc. Thesis, Babylon University,2002.

[4] Kuang-Chiung Chang, Cheng Wen, Ming-Feng Yeh, Ren-Guey Lee, " A Comparison Of Similarity Measures For Clustering Of QRS Complexes" , Received: August 12, 2005; Taoyuan, Taiwan, 2005.

[5] Silvia Selinski, " Similarity Measures For Clustering SNP And Epidemiological Data" , And The GENICA Network Interdisciplinary Study Group On Gene Environment Interaction And Breast Cancer In Germany, 2007.

[6] Anna Huang," Similarity Measures For Text Document Clustering", In The Proceedings Of The New Zealand Computer Science Research Student Conference, 2008.

[7] Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, Srinivas Mukkamala, Bernardete M. Ribeiro, " A Similarity Measure For Clustering And Its Applications ", By Icasa (Institute For Complex Additive Systems Analysis), A Division Of New Mexico Tech,2008.

[8] Mrutyunjaya Panda, Manas Ranjan Patra, " Some Clustering Algorithms To Enhance The Performance Of The Network Intrusion Detection System", Journal Of Theoretical And Applied Information Technology, 2008.

[9] Thomas Barrios, Rebecca Diamond, Guido W. Imbens, Michal Kolesar, "Clustering, Spatial Correlations And Randomization Inference", NBER Working Paper Series, 2010.

[10] K.P.N.V.Satya Sree , Dr.J V R Murthy, " Clustering Based On Cosine Similarity Measure", International Journal Of Engineering Science & Advanced Technology Volume-2, Issue-3, 508 – 512,2012.

[11] Saurabh Bhonde, P M Chawan, Prithviraj Chauhan, "Multi-Viewpoint Based Similarity Measure And Optimality Criteria For Document Clustering", International Journal Of Advanced Research In Computer Science And Software Engineering , Volume 2, Issue 6, June 2012 ISSN: 2277 128X

[12] R. Taylor, "Interpretation Of The Correlation Coefficient: A Basic Review" Journal Of Diagnostic Sonography, January 1990.

[13] J. Piantadosi, P. Howlett, J. Boland, "Matching The Grade Correlation Coefficient Using A Copula With Maximum Disorder", Journal Of Industrial And Management Optimization, Vol.3, No.2, Pp.305–312, 2007.

[14] Statsoft, Inc. "Semi-Partial (Or Part) Correlation", Electronic Statistics Textbook. Tulsa, OK: Statsoft, Accessed January 15, 2011.

[15] J. L. Rodgers And W. A. Nicewander. "Thirteen Ways To Look At The Correlation Coefficient". The American Statistician, 42(1):59–66, February 1988.

[16] W.Wang , Y.Zhang ,Yili , X.Zhang ,"The global fuzzy c-means clustering algorithm", In Proceedings of the World Congress on Intelligent Control and Automation, Vol. 1, 2006,pp. 3604–3607.

[17] T. Bäck, F. Hoffmeister and H. Schwefel, " A Survey of Evolution Strategies". In Proceedings of the Fourth International Conference on Genetic Algorithms and their Applications, pp. 2-9, 1991.

[18] A. Engelbrecht, "An Introduction ",Computational Intelligence, John Wiley and Sons, 2002.

[19] Y. Shi, R.C. Eberhart, "Parameter Selection in Particle Swarm Optimization", Evolutionary Programming VII: Proceedings of EP 98, 591-600, 1998.

[20] J. Kennedy, R. Eberhart, "Swarm Intelligence", Morgan Kaufmann, 2001.

[21] J.Kennedy, RC.Eberhart, "Particle Swarm Optimization", Proceedings of the IEEE International Joint Conference on Neural Networks, Vol. 4, pp 1942–1948,1995.

[22] R.C. Eberhart, Y.Shi, J.Kennedy "Swarm Intelligence", Morgan Kaufmann, 2002.

[23] D.Merwe, A.Engelbrecht , "Data Clustering using Particle Swarm Optimization". Congress on Evolutionary Computation , proceeding of IEEE , 2003.

[24] M.Omran, A. Engelbrecht and A. Salman, (2005). "Particle Swarm Optimization Method for Image Clustering", Pattern Recognition and Artificial Intelligence Vol. 19, No. 3, pp 297–321.

[25] M.T.Jones, "AI Application Programming", 2nd Ed. Hingham, Massachusetts: Charles River Media Inc, UK,2005.