A Study of Bio-inspired Algorithm to Data Clustering using Different Distance Measures

O.A. Mohamed Jafar
Research Scholar, Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi, Tamil Nadu, India

R. Sivakumar
Associate Professor, Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi, Tamil Nadu, India

ABSTRACT
Data mining is the process of extracting previously unknown and valid information from large databases. Clustering is an important data analysis and data mining method. It is the unsupervised classification of objects into clusters such that the objects from same cluster are similar and objects from different clusters are dissimilar. Data clustering is a difficult unsupervised learning problem because many factors such as distance measures, criterion functions, and initial conditions have come into play. Many algorithms have been proposed in literature. However, some traditional algorithms have drawbacks such as sensitive to initialization and easily trapped in local optima. Recently, bio-inspired algorithms such as ant colony algorithms (ACO) and particle swarm optimization algorithms (PSO) have found success in solving clustering problems. These algorithms have also been used in several other real-life applications. They are global optimization techniques. The distance based algorithms have been studied for the clustering problems. This paper provides a study of particle swarm optimization algorithm to data clustering using different distance measures including Euclidean, Manhattan and Chebyshev for well known real-life benchmark medical data sets and an artificially generated data set. The PSO-based clustering algorithm using Chebyshev distance measure is better fitness value than those of Euclidean and Manhattan distance measures.

Keywords
Data Mining, Data Clustering, Bio-inspired Algorithm, Particle Swarm Optimization, Distance Measures.

1. INTRODUCTION
Recent developments in information and computing technologies have resulted in computerizing many applications in various business areas. Data has played a vital role in many organizations. Every organization is accumulating a large volume of data and storing them as databases. This large amount of stored data contains valuable hidden knowledge, which could be used to improve the efficiency of business performance. Discovery of knowledge from this huge volume of data is indeed a challenge. There is a need for accessing the data, sharing the data and extracting useful information and knowledge from the data. The area of knowledge discovery in databases (KDD) has arisen over the last decade to address this challenge. Data mining is one of the steps in the process of knowledge discovery. It refers to extracting or mining useful information from large data sets. It is the process of extracting previously unknown, valid and potentially useful information from large databases. There are several data mining tasks including classification, regression, time series analysis, clustering, summarization, association rules and sequence discovery [1].

Data clustering is one of the important research areas in data mining. It is a popular unsupervised classification techniques which partitioning an unlabeled data set into groups of similar objects. The main aim of clustering is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters. It is an NP-hard problem of finding groups in data sets by minimizing some measure of dissimilarity. There are variety of clustering algorithms including K-means [2], K-medoids [3], BIRCH [4], DBSCAN [5], CURE [6], CHAMELEON [7], CACTUS [8], CLARANS [9], and K-Harmonic means [10]. Some of the algorithms have shortcomings such as initial point sensitivity, local optimal convergence, and global solutions of large problems cannot be found with reasonable amount of computational effort. In order to overcome these problems, many methods have been proposed.

Over the last decade, bio-inspired algorithms like PSO and ACO have found success in solving clustering problems [11]. In recent years, they have received special attention from the research community. Self organization, cooperation, communication, and flexibility are some of the important characteristics of bio-inspired algorithms. These algorithms have found to be robust in solving continuous optimization problems. In this paper, we have studied the performance of PSO algorithm to data clustering using different distance measures such as Euclidean, Manhattan and Chebyshev. This algorithm is experimented over four well-known real-world benchmark medical data sets and an artificially generated data set.

The rest of this paper is organized as follows: In section 2, data clustering and mathematical model of clustering problem are described. Section 3 introduces bio-inspired algorithms and fundamental principles of PSO. Data clustering using PSO algorithm is provided in section 4. A brief discussion of distance measures employed in PSO algorithm is given in section 5. The methodology is described in section 6. In section 7, the experimental results for the data sets used in this study are presented. Finally, section 8 concludes the paper and outlines the future work.

2. DATA CLUSTERING

2.1 Basic Concepts of Clustering

Clustering techniques partition the elements into clusters, so that elements within a cluster are similar to one another and dissimilar to elements in other clusters. Clustering is an example of unsupervised learning classification. The desirable properties of clustering algorithms are ability to deal with different data types, discovery of clusters with arbitrary shape, scalability, able to deal with noise and outliers, insensitive to order of input records, minimal requirements for domain knowledge to determine input parameters, incorporation of user-specified constraints, interpretability and usability [1].

Clustering is an important tool for a variety of applications in artificial intelligence, bioinformatics, biology, computer graphics, computer vision, data mining, data compression, earthquake studies, image processing, image segmentation, information retrieval, machine learning, marketing, medicine, network partitioning, object recognition, pattern recognition, spatial database analysis, routing, statistics, scheduling, vector quantization and web mining [12].

2.2 Major Clustering Methods

Clustering of data can be classified into partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, hard clustering methods and fuzzy clustering methods [1][12][13][14][15].

- ***Partitioning methods:*** The partition clustering techniques partition the database into predefined number of clusters. Given a database of 'n' objects, they attempt to determine 'k' groups, which satisfy the following requirements: (1) each object must belong to exactly one group, and (2) each group must contain at least one object. The popular algorithms of this type are K-means, K-mode, K-medoids, PAM, CLARA, and CLARANS.
- ***Hierarchical methods:*** Hierarchical clustering methods create a hierarchical decomposition of the objects. They can be either agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms start with each object forming a separate group. They successively merge the objects that are close to one another, until all the groups are merged into one or until a termination condition holds. Divisive algorithms begin with one object in a single cluster, then split the cluster into smaller groups until each object is in one cluster or a termination condition holds [16]. BIRCH, CURE, ROCK, and CHAMELEON are examples of these methods.
- ***Density-based methods:*** These algorithms cluster objects based on distance between objects. The data sets can be divided into several subsets according to the density of the data set points. The density is defined as the number of objects in a particular neighborhood of the data objects. DBSCAN, DENCLUE and OPTICS are typical density-based methods.
- ***Grid-based methods***: These methods quantize the object space into a finite number of cells that form a grid structure. They use a multi-resolution grid data structure. The famous algorithms of this kind are STING and CLIQUE.
- ***Model-based methods:*** These algorithms hypothesize a model for each of the clusters and find the best fit of the data to the given model. They can be either partitional or hierarchical depending on the structure or model. They follow the statistical modeling and neural-network based approach. EM and SOM are typical model-based methods.
- ***Hard and Fuzzy clustering methods:*** In hard clustering methods, each data point belongs to only one cluster. In fuzzy clustering methods, the data points can belong to more than one cluster and associated with each of the points, are membership grades. Membership degrees between zero and one are used. The degree of membership in this cluster depends on the closeness of the data objects to the cluster centers. Fuzzy-c means algorithm is the typical algorithm of this kind.

2.3 Mathematical Model of Clustering Problem

Given 'N' objects in R^m, allocate each object to one of 'K' clusters such that the sum of squared Euclidean distances between each object and the center of its belonging cluster for every such allocated object is minimized. The clustering problem can be mathematically described as follows [17]:

$$\text{Minimize } J(W,C) = \sum_{i}^{N}\sum_{j}^{K} w_{ij} \left\| x_i - c_j \right\|^2 \quad (1)$$

$$\text{where} \quad \sum_{j}^{K} w_{ij} = 1 \quad (2)$$

$$\text{and} \quad c_j = \frac{1}{n_j} \sum_{x_i \varepsilon C_j} x_i \quad (3)$$

K is the number of clusters;

N is the number of objects;

m is the number of object attributes;

x_i is the location of the i-th object;

c_j is the center of the j-th cluster;

$C = \{ C_1, \ldots, C_k \}$ denotes the set of K clusters;

$W = [w_{ij}]$ denotes the NxK 0-1 matrix; and

n_j denotes the number of objects belonging to cluster C_j.

3. BIO-INSPIRED ALGORITHM

Bio-inspired algorithms are the meta-heuristics that mimic the way nature performs for solving optimization problems. The field of bio-inspired computation covers many algorithmic approaches inspired by processes observed in nature such as

the evolution of species, emergent behavior of biological societies, and functioning of the vertebrate immune system. It includes well-known techniques such as ant colony optimization (ACO) [18] [19] artificial immune systems [20], artificial bee colony [21], evolutionary approaches [22], genetic algorithms [23] [24], particle swarm optimization (PSO) [25], tabu search [26] [27] and other methods to solve real-world problems. They overcome many limitations of traditional algorithms and have been widely accepted in science, engineering and business.

Recently bio-inspired algorithms have attracted many researchers. They have found numerous applications for solving problems from data mining [28], data clustering [29], classification [30], economic emissions load dispatch problem [31], travelling salesman problem [32] and others.

3.1 Fundamental Principles of PSO

PSO is a population-based stochastic optimization technique, which was proposed by Kennedy and Eberhart in 1995. It is inspired by the social behavior of animals such as a flock of birds, a school of fish, or a swarm of bees [33]. It is a global search procedure where the individuals, referred to as particles, are grouped into a swarm or population.

The working principle of the PSO may be described as follows: In PSO systems, the behaviors of animals are imitated by particles with certain positions and velocities in a multidimensional search space. Starting with a randomly initialized population, each particle in PSO flies through the searching space and remembers the best solution it has seen. Members of a swarm communicate good positions to each other and dynamically adjust their own position and velocity based on the good positions. The velocity adjustment is based upon the experiences and historical behaviors of the particles themselves as well as their neighbors. The performance of each particle is measured according to a predefined fitness function. The particles tend to fly towards better searching areas over the searching process. The positions of the particles are distinguished as personal best and global best.

For a D-dimensional search space, the position of the i-th particle is represented as $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, where d is the dimension number. A particle in a swarm is moving and has a velocity. The velocity of the i-th particle can be written as $V_i = (v_{i1}, v_{i2}, \ldots v_{id})$. At each iteration, the velocity and the position of a particle are updated based on its own previous best position and the global best position in the swarm. Each particle maintains a memory of its previous best position $P_i = (p_{i1}, p_{i2}, \ldots, p_{id})$. It is called as *pbest*. The best one among all the particles in the population is represented as $P_g = (p_{g1}, p_{g2}, \ldots, p_{gd})$. It is called as *gbest*. Positions and velocities are adjusted, and the function is evaluated with the new coordinates at each time step. The velocity and the position of the particle i are calculated using (4) and (5) respectively [25] [34].

$$v_{id}(t+1) = \omega \times v_{id}(t) + c_1 \times r_1 \times (p_{id} - x_{id}(t)) + c_2 \times r_2 \times (p_{gd} - x_{id}(t)) \quad (4)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (5)$$

where d - Dimension, d Є {1,2,…. D}

i - Index, i Є {1,2, … ,n}

n - Population size or number of particles in the swarm

c_1 - Cognitive component

c_2 - Social component

r_1 & r_2 - Uniformly generated random values in range [0,1]

v_{id} - Velocity of particle i on dimension d

x_i - Current position of the particle i on dimension d

p_{id} - Personal best or *pbest* position of particle i

p_{gd} - Global best or *gbest* position of the swarm

w - Inertia weight

4. DATA CLUSTERING USING PSO ALGORITHM

Data clustering using PSO algorithm was first introduced by Omran et al. 2002 [35]. This algorithm uses fixed number of clusters and uses PSO to search for the optimal centroids of these clusters. Van der Merwe and Engelbrecht [29] proposed a new approach for PSO-based clustering data. In this approach, all clusters will be continuously updated and moved toward the best cluster centroid. Esmin, Pereira and de Araujo [36] proposed a different approach to clustering data using PSO.

In the context of clustering, a single particle represents the N cluster centroid vectors. That is, each particle x is constructed as follows:

$x_i = (m_{i1}, m_{i2}, \ldots m_{ij} \ldots m_{iNc})$

where N_c is the number of clusters to be formed and m_{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.

Each particle is evaluated using the following equation:

$$J_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{\forall z_p \in C_{ij}} d(z_p, m_j) \Big/ |C_{ij}| \right]}{N_c} \quad (6)$$

where x_p denotes p-th data vector, $|C_{ij}|$ is the number of data of data vector belonging to the cluster C_{ij} and d is the distance between z_p and m_j.

Using the standard *gbest* PSO, data vectors can be clustered as follows:

1. Initialize each particle to contain N_c randomly selected cluster centroids.

2. For t = 1 to t_{max} do

a) For each particle i do

b) For each data vector z_p

i) Calculate distance $d(x_p, m_{ij})$ to all cluster centroids C_{ij}

ii) Assign z_p to cluster C_{ij} such that distance

$$d(z_p, m_{ij}) = \min_{\forall c = 1, \ldots, N_c} \{d(z_p, m_{ic})\}$$

iii) Calculate the fitness using equation (6)

iv) Update the global best and local best positions

v) Update the cluster centroids using the equations (4) and (5)

where t_{max} is the maximum number of iterations.

5. DISTANCE MEASURES

Cluster analysis assigns a set of 'n' objects to clusters on the basis of measurements of dissimilarity between the various objects. An important component of a clustering algorithm is the distance measure between data points. The distance measure will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another. The following are the important properties of distance measures [37] [38] [39] [40] [41] [42]:

Let d(a,b) denote the distance between points a and b.

1. Symmetry. The distance from a to b is the same as the distance from b to a. That is, d(a,b) = d(b,a).
2. Non-negativity. Distance is measured as a non-negative quantity. That is, $d(a,b) \geq 0$, for every a and b.
3. Identification. The distance between a and a is zero. That is, d(a,a) = 0, for every a.
4. Definiteness. If the distance between a and b is zero then a and b are the same. That is, d(a,b) = 0 only if a = b.
5. Triangle inequality. The length of one side of the triangle formed by any three points cannot be greater than the total length of the other two sides. That is, $d(a,c) \leq d(a,b) + d(b,c)$.

5.1 Euclidean Distance

Euclidean distance between two points is the shortest possible distance between the two points. It is also called Pythagorean metric since it is derived from the Pythagorean theorem. It is the commonly used distance measurement. It is invariant under orthogonal transformations of the variables. Many clustering algorithms have involved the use of Euclidean distances. One problem with the Euclidean distance measure is that it does not take the correlation between variables into account. Another drawback is that it does not work well for categorical variables [43].

It is also known as L_2 distance. It calculates root of square differences between the coordinates of a pair of objects. If a = (x_1, x_2) and b = (y_1, y_2) are two points, then the Euclidean distance between a and b is given by Eq. (7).

$$d(a,b) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (7)$$

If the points have 'n' dimensions such as a = $(x_1, x_2, \ldots x_n)$ and b = $(y_1, y_2, \ldots y_n)$ then the Euclidean distance between a and b is given Eq. (8).

$$d(a,b) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (8)$$

5.2 Manhattan Distance

Manhattan distance is also called city block distance or taxicab norm or rectilinear distance or L_1 distance. It computes the sum of absolute difference between the coordinates of a pair of objects [44]. If a = (x_1, x_2) and b = (y_1, y_2) are two points, then the Manhattan distance between a and b is given by Eq. (9).

$$d(a,b) = |x_1 - y_1| + |x_2 - y_2| \quad (9)$$

If the points have 'n' dimensions such as a = $(x_1, x_2, \ldots x_n)$ and b = $(y_1, y_2, \ldots y_n)$ then the Manhattan distance between a and b is given by Eq. (10).

$$d(a,b) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$$

$$= \sum_{i=1}^{n} |x_i - y_i| \quad (10)$$

5.3 Chebyshev Distance

It is also called L_α norm or minimax approximation or Chebyshev norm or Chessboard distance. It is named after Pafnuty Lvovich Chebyshev. It computes the absolute magnitude of the differences between coordinates of a pair of objects [45]. If a = (x_1, x_2) and b = (y_1, y_2) are two points, then the Chebyshev distance between a and b is given by Eq. (11).

$$d(a,b) = \text{Max} \{ |x_1 - y_1| , |x_2 - y_2| \} \quad (11)$$

If the points have 'n' dimensions such as a = $(x_1, x_2, \ldots x_n)$ and b = $(y_1, y_2, \ldots y_n)$ then the Chebyshev distance between a and b is given by Eq. (12).

$$d(a,b) = \text{Max} \{ |x_1 - y_1| , |x_2 - y_2| , \ldots , |x_n - y_n| \}$$

$$= \text{Max}_{\forall i=1,2,\ldots n} |x_i - y_i| \quad (12)$$

6. METHODOLOGY

The methodology used in this paper for data clustering is the PSO algorithm studied using different distance measures such as Euclidean, Manhattan and Chebyshev.

Case 1 Euclidean distance: The distance between the data vector x and centroid c is calculated by Eq. (13).

$$\sqrt{\sum_{i=1}^{n} (x_i - c_i)^2} \quad (13)$$

Case 2 Manhattan distance: The distance between the data vector x and centroid c is calculated by Eq. (14).

$$\sum_{i=1}^{n} |x_i - c_i| \quad (14)$$

Case 3 Chebyshev distance: The distance between the data vector x and centroid c is calculated by Eq. (15).

$$\text{Max}_{\forall i=1,2,\ldots n} |x_i - c_i| \quad (15)$$

The quality of PSO algorithm is measured according to the following criteria:

- The fitness of the particles is measured as the quantization error

$$J_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{\forall z_p \in C_{ij}} d(z_p, m_j) \Big/ |C_{ij}| \right]}{N_c}$$

where $d_{pj} = \| x_p - m_j \|$; $| C_{ij} |$ is the number of data vectors belonging to cluster C_{ij} ; and N_c is the number of clusters

- The maximum average distance (MAD) [46] is defined in Eq. (16)

$$d_{max}(Z_i, x_i) = \max \left\{ \sum_{\forall z_p \in C_{ij}} \frac{d(z_p, m_{ic})}{|C_{ij}|} \right\} \quad (16)$$

$$\forall c = 1, \ldots, N$$

7. EXPERIMENTAL RESULTS

The main objective of this study is to access the performance of the PSO algorithm to data clustering using different distance measures such as Euclidian, Manhattan and Chebyshev for well-known real-world benchmark data sets and an artificially generated data sets. The performance is measured by the fitness value and maximum average distance (MAD).

A. Parameter settings

Based on the experimental results the algorithm performs best under the following settings: The acceleration parameters (c_1 and c_2) are set to 2. The number of particles (p) is set to 10. V_{max} is set to ($X_{max} - X_{min}$) and V_{min} is set to $-(X_{max} - X_{min})$ [25]. The inertia weight (**ω**) is 0.9 → 0.4. That is, in general, **ω** decreases linearly from 0.9 to 0.4 throughout the search process [47]. **ω** is set to the following Eq. (17):

$$\omega = \omega_{max} - \frac{\omega_{max} - \omega_{min}}{I_{max}} * I ; \quad (17)$$

where ω_{max} and ω_{min} are the initial and final value of weighting coefficient respectively; I_{max} is the maximum number of iterations; I is the current iteration number, r_1 and r_2 are random values in the range [0,1].

B. Data sets

For evaluating the PSO based clustering algorithm, four well-known real-world benchmark data sets from the UCI machine learning repository and an artificially generated data set have been taken:

Data set 1: **Blood transfusion data set**, which consists of 748 instances and 2 different types characterized by 4 features. The features are recency – months since last donation, frequency – total number of donations, monetary – total blood donated in c.c., time – months since first donation.

Data set 2: **Pima Indians diabetes data set**: This data set is allocated to recognize diabetic patients. It consists of 768 instances which are classified into two classes consisting of 500 and 268 instances respectively. Each instance in this data set has 8 features, which are number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ml), body mass index (weight in kg / (height in m)^2, Diabetes pedigree function, and age.

Data set 3: **Liver disorder data set**, which consists of 345 objects and 2 different types characterized by 6 features including mcv mean corpuscular volume, alkphos alkaline phosphotase, sgpt alamine aminotransferase, sgot aspirate

aminotransferase, gammagt gamma-glutamyl transpeptidase, and drinks number of half-pint equivalents of alcoholic beverages drunk per day.

Data set 4: **Statlog (Heart) data set**, which consists of 270 objects. This data set contains 13 features, which are . age, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak=ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored by flourosopy, and thal.

Data set 5: **Artificial data set**: In this experimental data set, there are five classes and each class has 50 samples consisting of three features. Each feature of the class is distributed according to

Class1~Uniform(80,100),
Class1~Uniform(60,80),
Class1~Uniform(40,60),
Class1~Uniform(20,40), and
Class1 ~ Uniform(1,20).

Table 1 summarizes these five data sets. For each data set, it lists the number of instances, number of attributes and number of clusters.

C. Results

The algorithm tries to minimize the fitness value. The main purpose is to compare the performance of PSO algorithm with three different distance measures. The algorithm is implemented using Java. For our experimental tests, we used a PC Pentium IV (CPU 3.06 GHZ and 1.97 GB RAM) with the above parameter values by considering the maximum of 100 iterations, 10 particles and 10 independent test runs. The best, worst, average and standard deviation global fitness values and MAD are reported.

As shown in the tables 2 – 6 and Figs. 1 – 5, PSO based data clustering algorithm based on Chebyshev distance measure is better fitness value, standard deviation and MAD for all the data sets than those of Euclidean and Manhattan distances measures. Also the experimental results in Table 7 and Fig. 6 summarize the effect of varying the number of clusters for different distance measures for artificially generated data set. The fitness value should decrease when the number of clusters increase. It is also observed that PSO based data clustering algorithm based on Chebyshev distance measure shows minimum fitness value for varying number of clusters than those of other distance measures.

8. CONCLUSION

Clustering means the act of partitioning an unlabeled data set into groups of similar objects. Some traditional algorithms are sensitive to initialization and are easily trapped in local optima. On the other hand, the particle swarm optimization algorithm performs a globalized search in the entire solution space. Data clustering is a difficult problem because many factors such as distance measures, criterion functions and initial conditions have come into play. In this paper, particle swarm optimization algorithm is experimented with four well-known data sets Blood transfusion, Diabetes, Liver disorders, Statlog (Heart) and an artificially generated data set using different distance measures such as Euclidean, Manhattan and Chebyshev. This algorithm performs better fitness value for Chebyshev distance measure than the Euclidean and Manhattan distance measures for the data sets selected for our study. It is also observed that Manhattan distance measure performed very poorly in all the data sets. This distance measure is poor in handling high dimensional data.

Table 1 Description of data set used

Data set	No. of instances	No. of features	No. of clusters
Blood Transfusion	748	4	2
Diabetes	768	8	2
Liver Disorders	345	6	2
Statlog (Heart)	270	13	2
Artificial	250	3	5

Table 2 Best, worst, average, standard deviation of global fitness and maximum average distance (MAD) for 10 test runs of PSO algorithm on diabetes data set

Data set	Distance measure	Global fitness				MAD
		Best	Worst	Average	Std	
Diabetes	Euclidean	69.0695	109.4461	73.5667	6.2060	109.8465
	Manhattan	116.6343	209.0815	120.8809	12.0974	163.2181
	Chebyshev	**58.4621**	**95.0772**	**61.7756**	**5.9170**	**88.9208**

Table 3 Best, worst, average, standard deviation of global fitness and maximum average distance (MAD) for 10 test runs of PSO algorithm on blood transfusion data set

Data set	Distance measure	Global fitness				MAD
		Best	Worst	Average	Std	
Blood Transfusion	Euclidean	529.9936	788.9989	536.2758	27.1652	1045.9083
	Manhattan	545.0133	676.7063	561.9357	21.5307	1063.4851
	Chebyshev	**520.4634**	**583.9255**	**526.0786**	**13.8409**	**1030.1036**

Table 4 Best, worst, average, standard deviation of global fitness and maximum average distance (MAD) for 10 test runs of PSO algorithm on liver disorders data set

Data set	Distance measure	Global fitness				MAD
		Best	Worst	Average	Std	
Liver Disorder	Euclidean	32.2174	66.0195	38.0796	7.1726	36.8059
	Manhattan	59.7594	119.1086	84.8941	21.5288	78.5063
	Chebyshev	**25.4356**	**43.7139**	**27.3727**	**3.2734**	**32.8472**

Table 5 Best, worst, average, standard deviation of global fitness and maximum average distance (MAD) for 10 test runs of PSO algorithm on statlog (heart) data set

Data set	Distance measure	Global fitness				MAD
		Best	Worst	Average	Std	
Statlog (Heart)	Euclidean	40.3209	63.0872	42.9776	4.2062	44.3863
	Manhattan	71.9974	119.8488	75.5640	6.6570	73.7982
	Chebyshev	**32.9703**	**59.8969**	**33.5847**	**2.7279**	**36.7961**

Table 6 Best, worst, average, standard deviation of global fitness and maximum average distance (MAD) for 10 test runs of PSO algorithm on artificial data set

Data set	Distance measure	Global fitness				MAD
		Best	Worst	Average	Std	
Artificial	Euclidean	9.1511	16.2816	9.7812	1.2507	26.6418
	Manhattan	14.9483	24.4094	15.2170	1.2476	44.4797
	Chebyshev	**6.7890**	**10.2467**	**7.1944**	**0.6805**	**18.1727**

Table 7 Fitness value of different number of clusters

Data set	Distance measure	Number of clusters			
		2	3	4	5
Artificial	Euclidean	21.6930	17.7273	11.4437	9.1511
	Manhattan	37.4178	25.2146	18.8481	14.9483
	Chebyshev	**14.5203**	**9.6800**	**7.2602**	**6.7890**

Fig. 1 Comparison of global best fitness for 10 test runs in diabetes data set

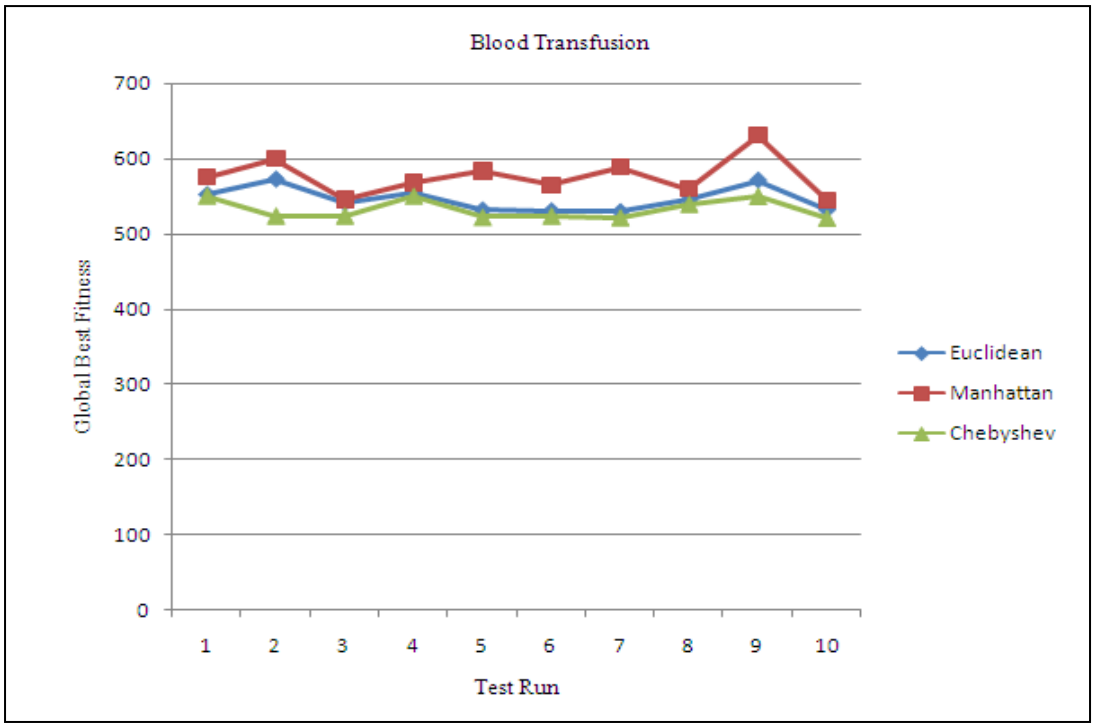

Fig. 2 Comparison of global best fitness for 10 test runs in blood transfusion data set

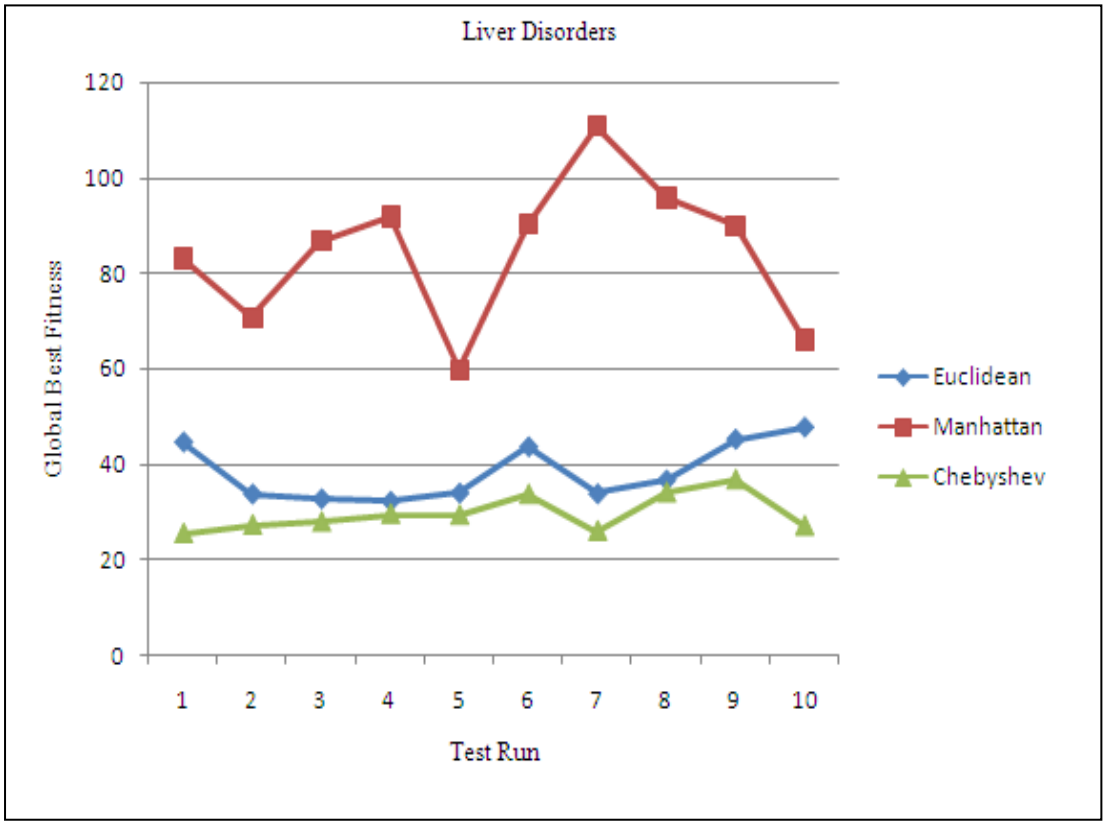

Fig. 3 Comparison of global best fitness for 10 test runs in liver disorder data set

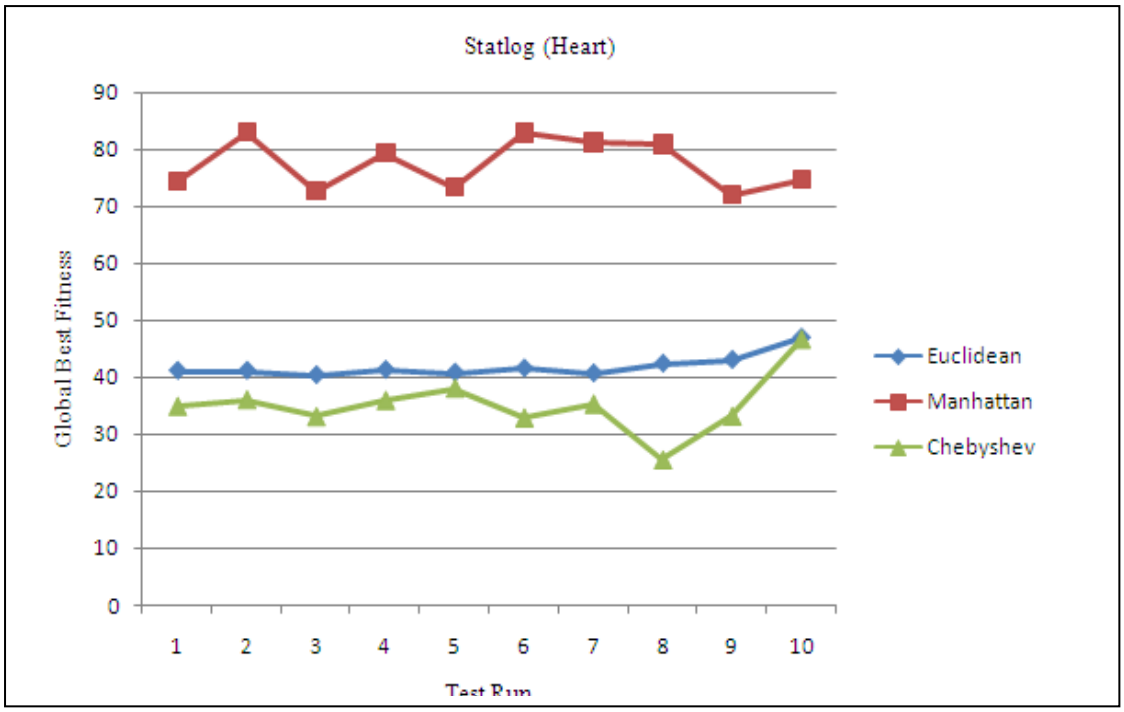

Fig. 4 Comparison of global best fitness for 10 test runs in statlog (heart) data set

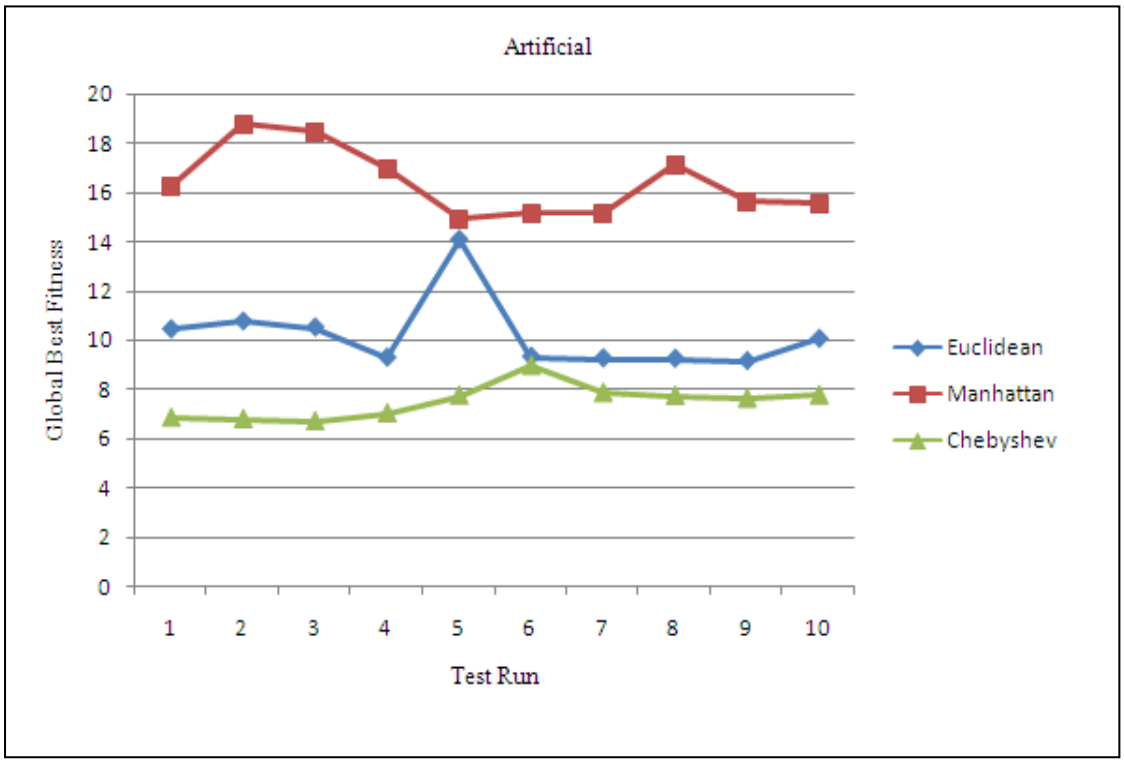

Fig. 5 Comparison of global best fitness for 10 test runs in artificial data set

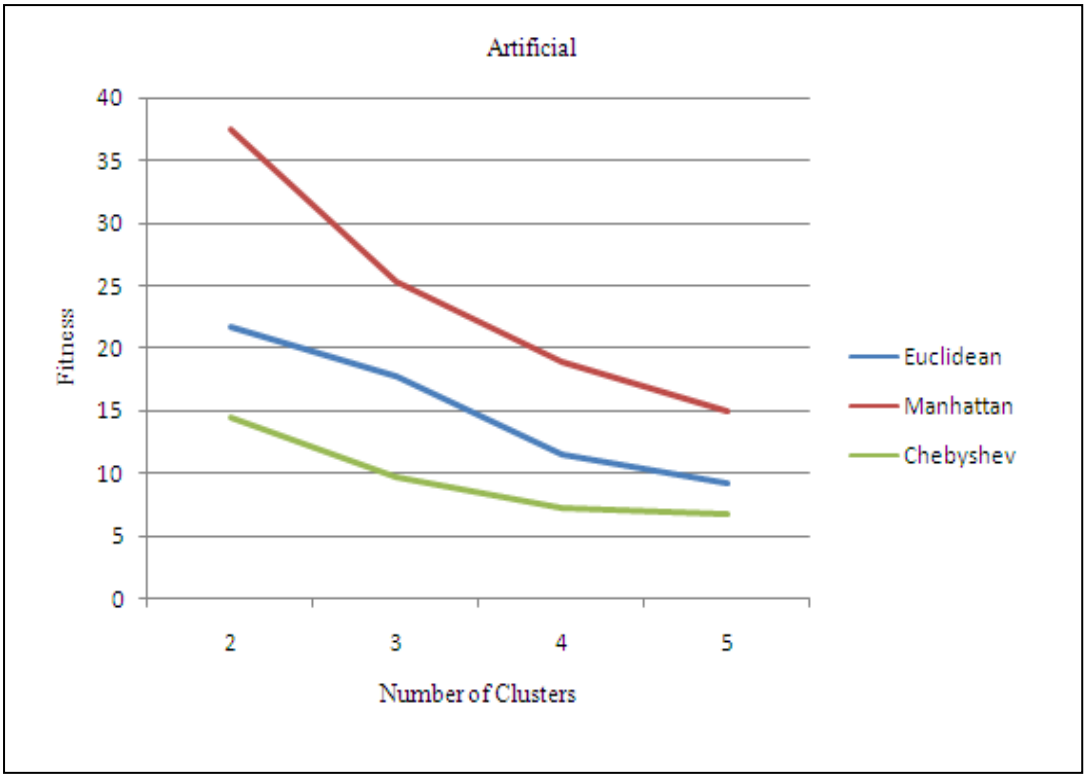

Fig. 6 Effect of the different number of clusters on the fitness value

9. REFERENCES

[1] Han, J., and Kamber. 2001. "Data mining: concepts and techniques", Morgan Kaufmann, San Francisco.

[2] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In 5th Berkeley symposium on mathematics, statistics and probability, pp. 281-296.

[3] Kaufman, L., and Russeeuw, P. 1990. "Finding groups in data: an introduction to cluster analysis", New York: John Wiley & Sons.

[4] Zhang, T., Raakrishanan, R., and Livny, M. 1996. "BIRCH: an efficient data clustering method for very large databases", In Proceedings ACM SIGMOD international conference on the management of data, pp. 103-114.

[5] Ester, M., Kriegel, H-P., Sander, J., and Xu X. 1996. "A density based algorithm for discovering clusters in large spatial databases with noise", In Simuoudis, E., Han, J., & Fayyard, U. editors, second international conference on knowledge discovery and data mining, pp. 226-231, AAAI press, Portland.

[6] Guha, S., Rastogi, R., and Shim, K. 1998. "CURE: an efficient clustering algorithm for large databases", In Proceedings ACM SIGMOD international conference on the management of data, pp. 73-84, Seatle, USA.

[7] Karypis, G., Han, E-H., and Kumar, V. 1999. "CHAMELEON: a hierarchical clustering algorithm using dynamic modeling", Computer, 32, pp. 32-68.

[8] Ganti, V., Gehrke, J., and Ramakrishnan, R. 1999. "CACTUS – clustering categorical data using summaries", In International conference on knowledge discovery and data mining, pp. 73-83, San Diego, USA.

[9] Ng, R., and Han, J. 2002. "CLARANS: a method for clustering objects for spatial data mining", IEEE Trans Knowl Data Eng, 14(5), pp. 1003-1016.

[10] Gungor, Z., and Unler, A. 2007. "K-harmonic means data clustering with simulated annealing heuristic", Applied mathematics and computation, 184(2), pp. 199-209.

[11] Bin, W., and Zhongzhi, S. 2001. "A clustering algorithm based on swarm intelligence", In Proceedings of the international conference on Info-tech and Info-net, Beijing, China, pp. 58-66.

[12] Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: a review. ACM Computing Surveys, 31(3), 264-323.

[13] Jain, A., and Dubes, R. 1998. "Algorithms for clustering data", Prentice Hall, New Jersey.

[14] Berkhin, P. 2002. "Survey clustering data mining techniques", Technical report, Accrue software, San Jose, California.

[15] Xu, R., and Wunsch II, D. 2005. "Survey of clustering algorithms", IEEE Transactions on Neural Networks, 16(3), 645-678.

[16] Ding, C., and He, X. 2002. "Cluster merging and splitting in hierarchical clustering algorithms", IEEE international conference, pp. 139-146.

[17] Yongguo Liu, Jun Peng, Kefei Chen, and Yi Zhang. 2006. "An improved hybrid genetic clustering algorithm", SETN 2006, LNAI 3955, pp. 192-202.

[18] Bonabeau, E., Dorigo, M., and Theraulaz, G. 1999. "Swarm intelligence: from natural to artificial systems", Oxford university press, Inc., New York.

[19] Dorigo, M., and Stutzle, T. 2004. "Ant colony optimization", MIT press, Cambridge, Massachusetts, London, England.

[20] de Castro, L.N., and Timmis, J. 2002. "Artificial Immune Systems: a new computational intelligence approach", Springer, Heidelberg.

[21] Zhang, C., Quyang, D., and Ning, J. 2010. "An artificial bee colony approach for clustering", Expert systems and applications, 37, pp. 4761-4767.

[22] Paterlini, S., and Minerva, T. 2003. "Evolutionary approaches for cluster analysis", In Bonarini. A., Musulli, F., Pasi, G., (Eds.) Soft computing applications, Springer-Verlag, Berlin, pp. 167-178.

[23] Goldberg, D.E. 1975. "Genetic algorithms in search, optimization and machine learning", Addison-Wesley, Reading, MA.

[24] Falkenauer, E. 1998. "Genetic algorithms and grouping problems", John Wiley and Sons, Chichester.

[25] Kennedy, J., and Eberhart, R.C. 1995. "Particle swarm optimization", In Proceedings of the IEEE international joint conference on neural networks, IJCNN 95, Piscataway, IEEE press, pp. 1942-1948.

[26] Al-Sultan, K.S. 1995. "A tabu search approach to the clustering problem", Pattern recognition, 28, pp. 1443-1451.

[27] Gendreau, M. 2003. "An introduction to tabu search", In Handbook of metaheuristics, Kochenberger, G., Glover, F., (Eds.), Dordrecht, Kluwer Academic Publishers.

[28] Sousa, T., Neves, A., and Silva, A. 2003. "Swarm optimization as a new tool for data mining",. In Proceedings of the 17th international symposium on parallel and distributed processing (IPDPS'03), pp. 48-53.

[29] Van der Merwe, D., and Engelbrecht, A. 2003. "Data clustering using particle swarm optimization", In Proceedings of IEEE congress on evolutionary computation (CEC 2003), Canbella, Australia, pp. 215-220.

[30] Liping Yan and Jianchao Zeng 2006. "Using particle swarm optimization and genetic programming to evolve classification rules", In Sixth world congress on intelligent control and automation (WCICA 2006), pp. 3415-3419.

[31] Apostolopoulos, T., and Vlachos, A. 2011. "Application of the firefly algorithm for solving the economic emissions load dispatch problem", International Journal of Combinatorics, 2011, pp. 1-23.

[32] Mustafa Servet Kiran, Hazim Iscan and Mesut Gunduz 2012. "The analysis of discrete artificial bee colony algorithm with neighborhood operator on travelling salesman problem", Neural computing and applications.

[33] Poli, R., Kennedy, J., and Blackwell, T. 2007. "Particle swarm optimization – an overview", Swarm intelligence, 1(1), pp. 33-57.

[34] Shi, Y., and Eberhart, R.C. 1998. "A modified particle swarm optimizer", In Proceedings of the IEEE congress on evolutionary computation (CEC 1998), Piscataway, NJ, pp. 69-73.

[35] Omran, M., Salman, A., and Engelbrecht, A. 2002. "Image classification using particle swarm optimization", In Wang L, Tan KC, Furukhashi T, Kim J-H, Yao X (Eds.), Proceedings of the fourth Asia-pacific conference on simulated evolution and learning (SEAL'02), IEEE press, Piscataway, pp. 370-374.

[36] Esmin, A.A.A., Pereira, D.L., and de Araujo, F. 2008. "Study of different approach to clustering data by using particle swarm optimization algorithm", In IEEE congress on evolutionary computation, CEC 2008, pp. 1817-1822.

[37] Sokal, R.R. 1977. "Clustering and classification: Background and current directions", Classification and clustering, Academic press, pp. 155-172.

[38] Mardia, K.V., Kent, J.T., and Bibby, J.M. 1979. "Multivariate analysis", Academic press.

[39] Seber, G.A.F. 1984. "Multivariate observations", Wiley.

[40] Mielke, P.W. 1985. "Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences", Journal of Atmospheric Sciences, 42, pp. 1209-1212.

[41] Krzanowski, W.J. 1988. "Principles of multivariate analysis: A user's perspective", Oxford science publications.

[42] Mimmack, Gillian M., Mason, Simon J., Galpin, and Jacquelin S. 2001. "Choice of distance matrices in cluster analysis: Defining regions", Journal of climate, 4(12), pp. 2790-2797.

[43] Ertoz, L., Steinbach, M., and Kumar, V. 2003. "Finding clusters of different sizes, shapes, densities in noisy high dimensional data", Proceedings of the third SIAM international conference on data mining (SDM 2003), volume 112, Proceedings in Applied mathematics, Society for industrial and applied mathematics.

[44] Berry, M.J.A., and Linoff, G.S. 2009. "Data mining techniques: For marketing, sales and customer relationship management", Second edition, Wiley.

[45] Bock, R.K., and Krischer, W. 1998. "The data analysis brief book", New York: Springer-Verlag.

[46] Omran, M.G.H. 2005. "A PSO-based clustering algorithm with application to unsupervised classification", University of Pretoria etd.

[47] Shi, Y., and Eberhart, R.C. 2002. "Empirical study of particle swarm optimization", In Proceedings of IEEE congress on evolutionary computation (CEC 1999), Washington D.C., pp. 1945-1949.