INTELLIGENT ANT BASED SOLUTION WITH DATA MINING ANALYSIS

Syeda Anum Rashid[1], Syed Jamal Hussain[2]
Department of Computer Science, University of Karachi, Pakistan
SyedaAnumRashid@hotmail.com, drsjhussain@yahoo.com

***Abstract*—This paper presents a study on how nature provides intelligent solutions to complex real-world problems.**
In this regard we study the intelligent behavior of swarming insect societies with the focus on ant colonies; examining how these almost blind species are capable of providing optimal solutions and how such solutions can be modeled into algorithms.
Further we have explored an interesting collaboration of data mining analysis with an ant based algorithm to prove that a nature inspired algorithm is able to perform data mining task successfully, this has been done with the help of a comparative experimental analysis.

***Keywords*—Intelligent solutions, swarm intelligence, stigmergy, cluster analysis**

I. INTRODUCTION

Algorithm development plays an integral role in the advancement of computer science. A major thrust in algorithm development is the design of algorithmic models to solve increasingly complex problems.

As most real world problems are non-linear and multimodal with uncertainty, they are often very challenging to solve, it is not possible to find the true global optimality with 100% certainty for a given problem. The increasing complexity of real-world problems motivates the researchers to search for efficient methods.

The traditional algorithms struggle in providing solutions to high complexity real world problem but nature provides many counter examples of biological systems exhibiting the required function. This is because the traditional algorithmic techniques lack the intelligent behavior that the biological systems possess. So in order to solve these computationally phrased problems optimally, algorithms need to provide intelligent solutions.

In this paper we examine the intelligent solutions provided by natural species and provide an experimental prove that they can perform better than traditional techniques.

The paper is structured as follows; in section II we explore intelligent solutions which are provided by the tiny little creations of Nature; the social insects. Then in section III we focus our study to a specific class of insects; ants, and in section IV an ant based clustering model is presented. While in section V a data mining analysis has been carried. It is a comparative experimental analysis, between ant based clustering and K-means clustering. Finally in section VI, on the basis of the experimental results obtained, the study is concluded.

II. INTELLIGENT SWARM BASED SOLUTIONS

Firstly **what are intelligent solutions**? Intelligent solutions are ones that can provide optimal results for real world complex problems, they are inspired by the intelligent behavior of natural species.

How can an algorithm provide intelligent solution? To determine what general qualities an intelligent algorithm must possess. Consider the problems of **Face recognition** and **Speech recognition.** Both problems are of current and trending interest, these are difficult to solve via computers but obviously rather easy for humans and even animals. Why natural brains find such problems easy to process? This is because the natural brain possess the following characteristics:

- Ability to extract models from examples (**learning**)
- Ability to deal with dynamic environments (**adaptability**)
- Ability to deal with noisy, incomplete or inconsistent data (**robustness**)
- Ability to provide the answer in a reasonable amount of time (**efficiency**).

These characteristics also define the general qualities of intelligent algorithms, so algorithms must mimic the above basic traits of the natural brain to work intelligently!

Algorithms and techniques mimicking the different behavioral pattern displayed by various natural species are classified under the umbrella of **Nature Inspired Algorithms (NIA)**. These algorithms are meta-heuristic algorithms [1]; higher level heuristic algorithms (here *heuristic* means "to find" or "to discover by trial and error). In general these nature-inspired meta-heuristic algorithms can be categorized as:

- Biological systems based algorithms
  - Swarm intelligence (SI) based
  - Non Swarm intelligence (SI) based
- Physical and chemical systems based algorithms

By far the majority of nature-inspired algorithms are based on some successful characteristics of biological system. Therefore, the largest fraction of nature-inspired algorithms is biology-inspired [2]. Same is our topic of concern with the study being focused on social insects, so we will only be discussing those Biological systems based algorithms which are Swarm Intelligent (SI) based.

But **what is Swarm intelligence**? It is the Collective intelligent behavior of decentralized and self-organized systems [2], [3]. While each individual of the systems may be considered unintelligent, but the whole system of multiple agents may show collective intelligence.

Similarly individuals in social insect societies of ants, bees, termites, wasps and all others, can be considered unintelligent but when they work as a group they exhibit highly structured self-organization and hence show collective intelligence. So, social insect practice swarm intelligence and the algorithms formulated from the practices of these insects are termed as Swarm intelligence (SI) based algorithms. Few popular SI based algorithms are listed in table 1:

Algorithm	Author	Reference
Particle Swarm optimization	Yang et al.	[4]
Ant colony optimization	Marco Dorigo	[5]
Bat algorithm	Xin-She Yang	[7]
Cuckoo search	Yang and Deb	[8]
Firefly algorithm	Xin-She Yang	[9]
Flower pollination	Xin-She Yang	[10]

Table 1 SI based algorithms [2]

All the above mentioned algorithms are based on swarm intelligence characteristic and have applications in various fields, some of which are listed below:

Algorithm	Characteristic	Application
Particle Swarm optimization	Uses the swarming behavior of fish and birds	Multi-dimensional searches, neural networks, antenna design, electromagnetics [4]
Ant colony optimization	Uses the capability of ants to find the shortest path	Bus routes, delivery routes, machine scheduling, telecommunication networks [6]
Bat algorithm	Uses the echolocation of foraging bats, sonar pulses to detect prey and avoid obstacles	Shows good results when dealing with lower-dimensional optimization problems [7]
Cuckoo search	Uses the brooding parasitism (aggressive reproduction strategy) of some cuckoo species	Applied successfully to determine the quality or fitness of a solution for a maximization problem [8]
Firefly algorithm	Uses the flashing behavior of swarming fireflies	It has been used to carry out various design optimization tasks [9]
Flower pollination	Uses the phenomenon of pollen transfer	Can outperform both Genetic and Particle Swarm Optimization [10]

Table 2 Characteristics of SI Based Algorithms

III. SWARMING BEHVIOR OF ANTS

Without a doubt the most successful specie on earth today, is that of ants in fact they have been so, for the past 100 million years. The study of the behavior of ant colonies and of their self-organizing capacities is interesting for computer scientists because it provides models of distributed organization which are useful to solve difficult optimization and distributed control problems.

Primarily ants in an ant colony use the stigmergic approach to co-ordinate their activities. Stigmergy is a particular form of indirect communication used by social insects to coordinate. During the stigmergic approach of ants the basic behaviors observed are:

A Foraging

Ants have a trail-laying/trail-following behavior when foraging; searching for food [11]: individual ants deposit pheromone while walking and foragers follow pheromone (a chemical substance produced and released into the environment by an animal, affecting the behavior or physiology of others of its species) trails with some probability. This behavior can explain how ants find the shortest path between their nest and a food source. Consider figure 1-a where ants are coming back and forth from their nest to food source.

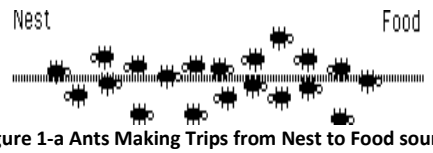

Figure 1-a Ants Making Trips from Nest to Food source

In figure 1-b an obstacle is placed in the path from ant nest to food source, this has created two different length paths form nest to food source.

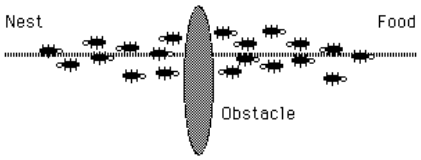

Figure 1-b an Obstacle in Ant Route

Ants will select randomly one of the two paths, with equal probability as seen in figure 1-c.

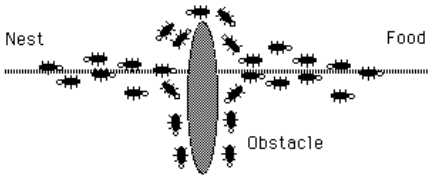

Figure 1-c Obstacle Creating Two Paths

Those ants will return earlier to the nest, who took the shortest route twice (to go from the nest to the source and to return to the nest), so that more pheromone is present on the shorter path than on the longer path immediately after these ants have returned, stimulating other nest mates to choose the shorter path as well.

Eventually the greater amount of pheromone on one path stimulates more ants to choose it, and so on. This autocatalytic process will lead the ant colony to converge towards the use of only one of the two paths, which is the shorter path in this case. As shown in figure 1-d.

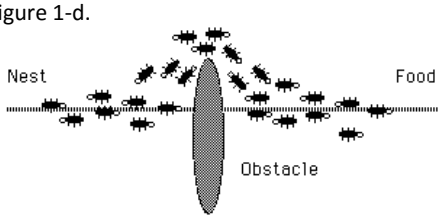

Figure 1-d Ants Selecting the Shorter Path

This has been called differential length effect [12] and explains how ants in the long run end up choosing the shorter of the two paths without using any global knowledge about their environment. It is also interesting to note that in some ant species the amount of pheromone deposited is proportional to the quality of the food source found: paths that lead to better food sources receive a higher amount of pheromone.

Differential length effect and pheromone based autocatalysis are at the earth of some successful ant algorithms for discrete optimization, in which an artificial pheromone plays the role of stigmergic variable.

B Division of labor

Division of labor is an important and widespread feature of life in ant colonies, and in social insects in general. Social insects are all characterized by one fundamental type of division of labor, reproductive division of labor. Consider figure 2, which illustrates four major casts of ants.

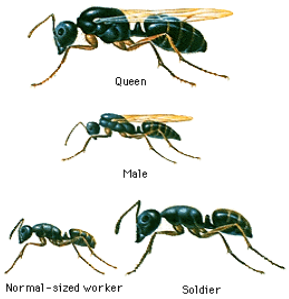

Figure 2 Different Ant Casts

Beyond this primary form of division of labor between reproductive and worker castes, there often exists a further division of labor among workers, who tend to perform specific tasks for some amount of time, rather than to be generalists who perform various tasks all the time. Workers are divided into

- Age or
- Morphological sub-castes.

Age sub-castes correspond to individuals of the same age that tend to perform identical tasks, this phenomenon is called temporal polyethism. In some species, workers can have different morphologies, workers that belong to different morphological castes tend to perform different tasks.

But even within an age or morphological caste, there may be differences among individuals in the frequency and sequence of task performance, one may therefore speak of behavioral castes, to describe groups of individuals that perform the same set of tasks in a given period.

One of the most striking aspects of division of labor is plasticity, a property achieved through the workers' behavioral flexibility: the ratios of workers performing the different tasks that maintain the colony's viability and reproductive success can vary (i.e., workers switch tasks) in response to internal perturbations or external challenges.

It is amazing to observe how this flexibility is implemented at the level of individual workers, which do not possess any global representation of the colony's needs. This technique of labor division can provide solutions to complex task allocation problems.

C Cemetery formation

Intensive experiments reported that many species along with ants actually organize a cemetery [12]. The phenomenon that is observed in experiments is the aggregation of dead bodies by worker ants. If dead bodies, or more precisely items belonging to dead bodies, are randomly distributed in space at the beginning of the experiment, the workers will form clusters within a few hours (figure 3). This figure shows four successive pictures of circular arena. After 3, 6 and 36 hours.

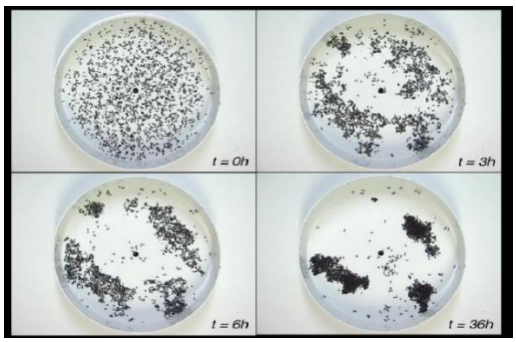

Figure 3 Forming of Clusters in Real Ants

If the experimental arena is not sufficiently large, or if it contains spatial heterogeneities, the clusters will be formed along the borders of the arena or more generally along the heterogeneities.

The basic mechanism underlying this type of aggregation phenomenon is an attraction between dead items mediated by the ant workers: small clusters of items grow by attracting workers to deposit more items.

It is this positive feedback that leads to the formation of larger and larger clusters. In this case it is therefore the distribution of the clusters in the environment that plays the role of stigmergic variable.

IV. AN ANT BASED CLUSTERING MODEL

Deneubourg et al. [13] have proposed a model relying on biologically plausible assumptions to account for the phenomenon mentioned in section III-C of dead body clustering or cemetery organization in ants.

The model, called basic model (BM), relies on the general idea that isolated items should be picked-up and dropped at some other location where more items of that type are present. Let us assume that there is only one type of item in the environment. The probability for a randomly moving ant (that is currently not carrying an item) to pick-up an item is given by

$$P_p = \left(\frac{k_1}{k_1 + f}\right)^2 \qquad (1)$$

where f is the perceived fraction of items in the neighborhood of the ant and k_1 is the threshold constant: for $f \ll k_1$, P_p is close to 1 (i.e., the probability of picking-up an item is high when there are not many items in the neighborhood), and P_p is close to 0 if $f \gg k_1$ (i.e., items are unlikely to be removed from dense clusters). The probability P_d for a randomly moving loaded ant to deposit an item is given by

$$P_d = \left(\frac{f}{k_2 + f}\right)^2 \qquad (2)$$

Where k_2 is another threshold constant: for $f \ll k_2$, P_d is close to 0, whereas for $f \gg k_2$, P_d is close to 1.

As expected, the pick-up and deposit behaviors obey roughly opposite rules. The question is now to define how f is evaluated. Deneubourg et al. [13], having in mind a robotic implementation, moved away from biological plausibility and assumed that f is computed using a short-term memory that each ant possesses: an ant keeps track of the last T time units, and f is simply the number N of items encountered during these last T time units divided by the largest possible number of items that can be encountered during T time units.

If one assumes that only zero or one object can be found within a time unit, then f = N/T. Figure below shows a simulation of this model: small evenly spaced clusters emerge within a relatively short time and then merge into fewer larger clusters (figure 4).

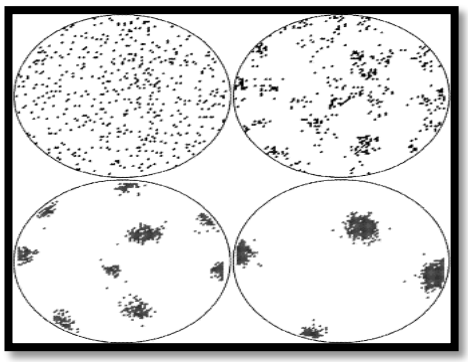

Figure 4 Computer Simulation of Clustering Model

The figure shows three successive pictures of simulated circular arena. From left to right and from up to down: the initial state with 5000 items placed randomly in the arena, the arena at t=50,000, t=

100,000 and t=5 000 000. Parameters: T= 50, k_1= 0.1, k_2= 0.3, 10 ants.

Lumer and Faieta [14] have generalized Deneubourg et al.'s BM [13], to apply it to exploratory data analysis. The idea here is to define a "dissimilarity" d (or distance) between objects in the space of object attributes. This algorithm by Lumer and Faieta [14] (referred hereafter as LF algorithm) works as follows.

Let us assume that the number of dimensions, m = 2; instead of embedding the set of objects into $\mathbb{R}^2$, the LF algorithm approximates this embedding by considering a grid, i.e., a subspace of $\mathbb{Z}^2$. Ants can directly perceive a surrounding region of area $s^2 - 1$ (a square $Neigh_{s*s}$ of s*s sites surrounding site r).

Let d (o_i, o_j) be the distance between two objects o_i and o_j in the space of attributes. Let us also assume that an ant is located at site r and at time t, and finds an object o_i at that site. The local density of objects similar to type o_i at site r is given by

$$f(o_i) = \max\left\{\frac{1}{s^2}\sum_{o_j \in Neigh(r)}\left(1 - \frac{d(o_i,o_j)}{\alpha}\right)\right\} \qquad (3)$$

Here $f(o_i)$ is a measure of the average similarity of object o_i with the other objects o_j present in its neighborhood: this expression replaces the fraction f of similar objects of the BM.

The parameter α defines the scale for dissimilarity: its value is important for it determines when two items should or should not be located next to each other. Lumer and Faieta [14] define picking-up and dropping probabilities as follows:

$$P_p(o_i) = \left(\frac{k_1}{k_1 + f(o_i)}\right)^2 \qquad (4.1)$$

$$P_d(o_i) = \begin{cases} 2f(o_i) & if\ f(o_i) < k_2 \\ 1 & if\ f(o_i) \geq k_2 \end{cases} \qquad (4.2)$$

Where k_1 and k_2 are two constants that play a role similar to k_1 and k_2 in the BM.

The LF algorithm has been implemented and successfully extended in cluster analysis. We have also implemented the algorithm on matlab, a brief description of it is discussed below:

The algorithm is designed with the help of equations 4.1 and 4.2 and the following parameters and key variables:

- lambda: the measure of average similarity (previously referred to as $f(o_i)$)
- pickupP: the probability of picking up a pattern, of an ant
- dropP: the probability of putting down a picked up pattern, of an ant
- alpha: dissimilarity scale; given by user at runtime
- Number of Ants: used to store the number of ants
- Pick Up Gamma: threshold constant k_1; given by user at runtime [Equation 4.1]
- Drop Gamma: threshold constant k_2; given by user at runtime [Equation 4.2]
- Maximum Iteration: the maximum number till the clustering will be executed continuously
- Direction Probability: used to store the direction probability of ants
- Step Size: used to store the step size of ants
- Neighbor Size: used to store the neighborhood size
- Number of Patterns: is the count of patterns that is to be randomly clustered for clustering
- Number of Clusters: it is the count of the types of patterns
- Number of Features: number of features associated with the patterns; given by user at runtime

The general description of the algorithm is:

Initialization phase

- Values of Number of Patterns, Number of Clusters, Number of Features, alpha, Number of Ants, Pick Up Gamma, Drop Gamma, Maximum Iteration, Direction Probability, Step Size and Neighbor Size are initialized; assigning each ant a direction.
- To each ant, a pattern load and a location field is assigned

Clustering phase

for 1 to maximum Iteration
  for 1 to number of Ants
    Move each ant in a specific direction
    Setup neighborhood of each ant
    Pick up pattern by ant if ant is empty with the probability
    $pickupP = \left(\frac{gamaPick}{gamaPick + lambda}\right)^2$ [Equation 4.1]
    Where lambda is
    $lambda = max\left\{\frac{1}{t} * sum\left(1 - \frac{d}{alpha}\right)\right\}$
    Drop pattern by ant if is not empty and location is available
    The drop probability is given by
    $DropP = \begin{cases} 2 * lambda\ if\ lambda < gamaDrop \\ 1 \quad if\ lambda \geq gamaDrop \end{cases}$ [Equation 4.2]
    Where again lambda is
    $lambda = max\left\{\frac{1}{t} * sum\left(1 - \frac{d}{alpha}\right)\right\}$
  end
end

Figure 5 presents the GUI of the LF algorithm based clustering program, in figure 5-a, two types of patterns are randomly distributed, since 2 is the input in the number of clusters field. While in figure 5-b ants (diamond shaped) are clustering the patterns.

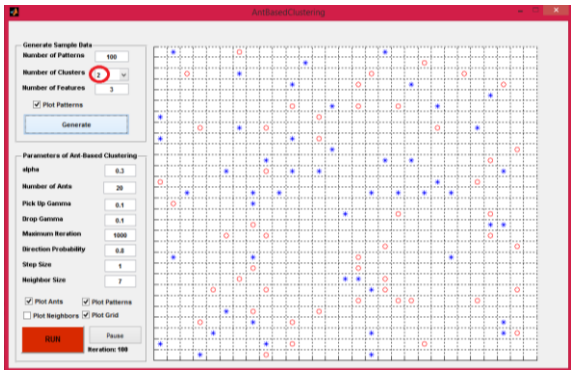

Figure 5-a Pattern Plot

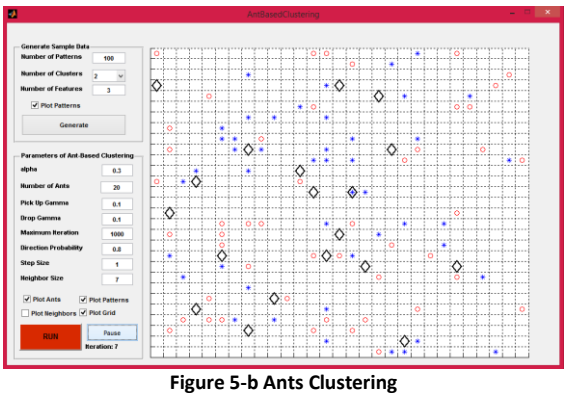

Figure 5-b Ants Clustering

The performance of this clustering program can be summed up in the table 3. Here number of cluster are the type of patterns.

NUMBER OF CLUSTERS	ITERATIONS TAKEN TO CLUSTER	NUMBER OF ANTS	alpha (α)
2	318	20	0.3
3	385	20	0.3
4	610	20	0.3
5	966	20	0.3

Table 3 Performance of Ant Based Clustering

V. A COMPARATIVE DATA MINING ANALYSIS BETWEEN ANT BASED CLUSTERING AND K-MEANS ALGORITHM

The traditional clustering algorithms such as K-means, K- mediods, Fuzzy C-means cannot be applied to all applications as these algorithms have a few drawbacks like, initial partition selection and local optima convergence. Therefore approaches similar to ant based clustering have been introduced to overcome these drawbacks.

Cluster Analysis is actually a popular data analysis or a data mining technique as shown in figure 6. So we present a data mining analysis. The analysis has been carried out on a telecom company's customer data set, comprising 2669 cases of customers.

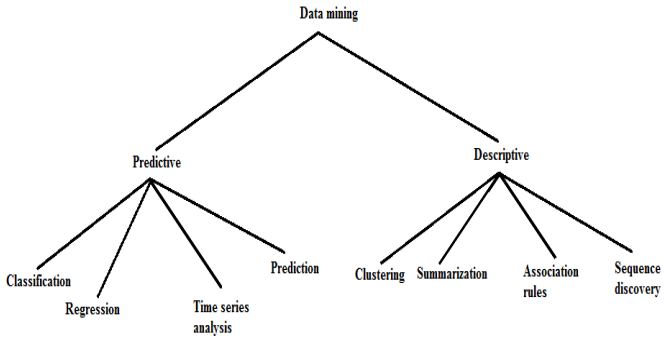

Figure 6 Data Mining

The data attributes used in the experiment are as follows:

VARIABLE NAME	DESCRIPTION
Regular_dur	Minutes of call in regular time
Discount _dur	Minutes of call in discount time
Local_ dur	Minutes of local call
Domestic_dur	Minutes of domestic call
Svc_sms	Times of short message service
Svc_type	Number of service types
Svc_time	Number of service times
Age	Age Customer age
Gender	Customer gender
Balance	Balance of customer account
Arrearage_time	Times of arrearage
ARPU	Average Revenue Per User
Churn	Customer is churning or not

Table 4 Data Attributes of Clustering Experiment

Parameters of Ant-Cluster algorithm are set as follow in this experiment. Similarity coefficient is α=12~14, the maximum number of iterations is 8000, the number of ants in each population is 100, threshold constants (Pick Up Gamma and Drop Gamma) are k_1=0.l and k_2=0.15. Clustering result is illustrated in figure 7.

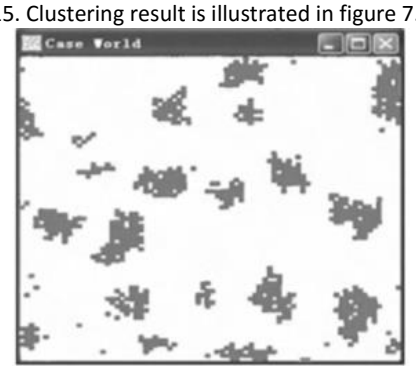

Figure 7 Result of Customer Segmentation by Ant Cluster Algorithm

In figure 7, each cluster figures one customer cluster. Objects in a cluster have some common characteristics and these characteristics can be obtained by comparing distribution of an attribute value in the whole data set with the one in a certain cluster.

For example, figure 8 and 9 represent distribution of domestic call time attribute in the whole customer data set and in a certain customer cluster respectively.

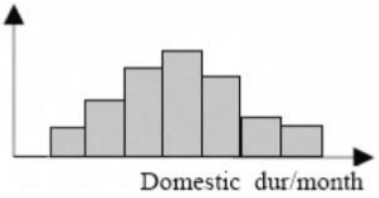

Figure 8 Distribution of Domestic Call Time in All Customer Data Set

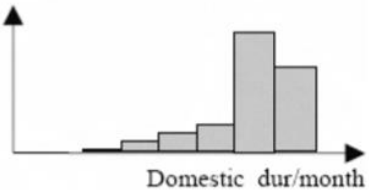

Figure 9 Distribution of Domestic Call Time in Certain Customer Cluster

As is shown in figure 9, domestic call time in this cluster is longer than that in all customer data set shown in figure 8. Therefore, we can draw a conclusion that higher domestic call time is one of characteristics of this cluster. Appropriate marketing strategies should be made according to this result.

Figure 10 shows the average value of Domestic_dur in each cluster obtained with Ant-based clustering algorithm.

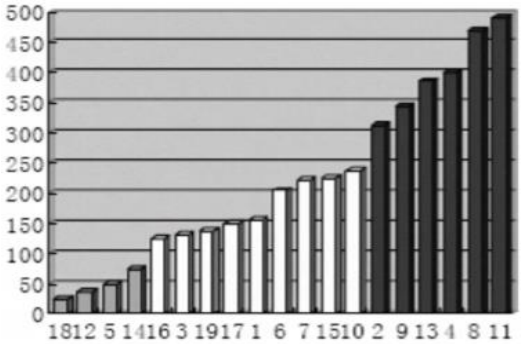

Figure 10 Result of Ant based Clustering Algorithm

For comparing, we implemented the k-means algorithm. The average value of Domestic_dur in each cluster obtained with k-means algorithm (k=19) is shown in figure 11.

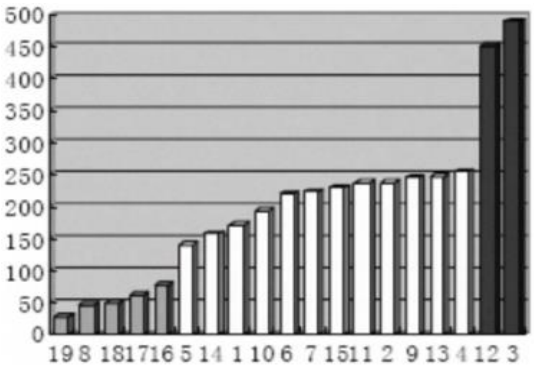

Figure 11 Result of K-means Algorithm

In figure 10, the Ant-Cluster algorithm has obtained six clusters which average value of Domestic_dur are more than 300, namely 2, 9, 13, 4, 8, and 11. The k-means algorithm has only discovered two clusters which average value of Domestic_dur are more than 300, namely 2 and 3.

We can conclude that Ant based clustering algorithm is more effective than k-means algorithm for discovering clusters which have distinct characteristics when the numbers of clusters are same or similar.

VI. CONCLUSION

We went through the process of extracting solutions from a natural specie; ant and formulated into an algorithm, successfully implementing it to perform a data mining or more precisely clustering analysis.

The result of our analysis clearly proves that nature inspired algorithms are intelligent enough to solve real world complex problems. Moreover the results are better than the traditional algorithms.

VII. FUTURE WORK

The above discussed and other similar natural techniques of computing yield special attention because of their performance as they can solve real world problems with efficiency and accuracy. Currently nature based algorithms are aiding the evolving process of fields such as Digital Image Processing, Business Intelligence, Knowledge Discovery, Big Data Analysis, Ubiquitous Computing, Telecommunication and the list continues.

Based on the potential of this extremely rich domain, we can say that natural computing will soon become the most used from of computing.

REFERENCES

[1]. Xin-She Yang**, "Nature-Inspired Optimization Algorithms**", School of Science and Technology**,** Middlesex University Press, London, 2014.

[2]. Iztok Fister Jr., Xin-She Yang, Iztok Fister, Janez Brest, Duˇsan Fister, "**A Brief Review of Nature-Inspired Algorithms for Optimization",** ELEKTROTEHNIˇSKI VESTNIK 80(3): 1–7, 2013. English edition.

[3]. Arne Brutschy, Alexander Scheidler, Daniel Merkle, and Martin Middendorf, "**Learning from House-Hunting Ants: Collective Decision-Making in Organic Computing Systems",** Ant Colony Optimization and Swarm Intelligence, 6th International Conference, ANTS 2008 Brussels, Belgium, September 22-24, 2008 Proceedings.

[4]. James Kennedy and Russell Eberhart, "**Particle Swarm Optimization"**, Washington, Purdue School of Engineering and Technology Indianapolis, 1995.

[5]. Marco Dorigo, "**Optimization, learning and natural algorithms**", Ph. D. Thesis, Polytechnic of Milano, Italy, 1992.

[6]. M. Dorigo & T. Stützle, "**the Ant Colony Optimization Meta-heuristic: Algorithms, Applications, and Advances**", Handbook of Meta-heuristics, 2002.

[7]. IztokFister Jr., Duˇsan Fister, and Xin-She Yang, "**A Hybrid Bat Algorithm"**, arXiv:1303.6310v3 [cs.NE], 5 Jun 2013.

[8]. Xin-She Yang,Suash Deb, "**Cuckoo Search via L´evy Flights",** arXiv: 1003.1594v1 [math.OC], 8 March 2010.

[9]. Xin-She Yang**, "Firefly Algorithm, Stochastic Test Functions and Design Optimization",** Department of Engineering, University of Cambridge, arXiv: 1003.1409v1 [math.OC], 6 March 2010.

[10]. Xin-She Yang, "**Flower Pollination Algorithm for Global Optimization",** Department of Engineering, University of Cambridge**,** 2013.

[11]. B. Hölldobler, E.O. Wilson,"**the Ants",** Springer, Berlin, 1990

[12]. Marco Dorigo, Eric Bonabeau, Guy Theraulaz, "**Ant algorithms and stigmergy",** Future Generation Computer Systems 16 (2000) 851–871.

[13]. J.-L. Deneubourg, S. Goss, N. Franks, A. Sendova-hanks, C. Detrain, L. Chrétien, "**The dynamics of collective sorting: robot-like ants and ant-like robots"**, in: J.-A. Meyer, S.W. Wilson (Eds.), Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animats, MIT Press/Bradford Books, Cambridge, MA, 1991, pp. 356–363.

[14]. E. Lumer and B. Faieta, **"Diversity and adaptation in populations of clustering ants"**, Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats, Vol. 3, MIT Press/Bradford Books, Cambridge, MA, 1994, pp. 501–508.