ANALYZING AND OPTIMIZING ANT-CLUSTERING ALGORITHM BY USING NUMERICAL METHODS FOR EFFICIENT DATA MINING

Md. Asikur Rahman[1], Md. Mustafizur Rahman[2], Md. Mustafa Kamal Bhuiyan[3], and S. M. Shahnewaz [4]

[1]Department of Computer Science, Memorial University of Newfoundland, St. John's, Canada
asikur.rahman@mun.ca
[2]Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Canada
m.rahman@mun.ca
[3]Department of Computer Science, Memorial University of Newfoundland, St. John's, Canada
mmkb65@mun.ca
[4]Department of Electrical and Computer Engineering (ECE), University of Calgary, Calgary, Canada
smshahne@ucalgary.ca

ABSTRACT

Clustering analysis is an important function of data mining. There are various clustering methods in Data Mining. Based on these methods various clustering algorithms are developed. Ant-clustering algorithm is one of such approaches that perform cluster analysis based on "Swarm Intelligence'. Existing ant-clustering algorithm uses two user defined parameters to calculate the picking-up probability and dropping probability those are used to form the cluster. But, use of user defined parameters may lead to form an inaccurate cluster. It is difficult to anticipate about the value of the user defined parameters in advance to form the cluster because of the diversified characteristics of the dataset. In this paper, we have analyzed the existing ant-clustering algorithm and then numerical analysis method of linear equation is proposed based on the characteristics of the dataset that does not need any user defined parameters to form the clusters. Results of numerical experiments on synthetic datasets demonstrate the effectiveness of the proposed method.

KEYWORDS

Ant-Clustering Algorithm, Swarm Intelligence, Numerical Method, Linear Equations

1. INTRODUCTION

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics.

DOI : 10.5121/ijdkp.2012.2501

Clustering is one of the widely used knowledge discovery techniques to reveal structures in a dataset that can be extremely useful to the analyst [1]. It is a kind of unsupervised learning method that groups a set of data objects into clusters. In these clusters, data objects are similar to one another within the same cluster and are dissimilar to the objects in other clusters [2]. Major clustering methods are classified into five categories, i.e.

a) partitioning methods
b) hierarchical methods
c) density-based methods
d) grid-based methods, and
e) model-based methods.

Some clustering algorithms integrate the ideas of several clustering methods. But each kind of clustering methods has its own limitation. For example, many algorithms (like K-means and ISODATA used in the following) require that an initial partition is given as input before the data can be processed [3], [4]. This is one major drawback for these methods, and it is important to notice that ant-based approaches to clustering do not require such an initial partitioning. Although the ant-clustering algorithm do not require initial partitioning, but it is not out of limitations [2].

In "Ant-Cluster algorithm", multi-population of ants with different moving speed is introduced and outlier objects are processed properly. Ants are the agents which can pick up or drop out the data objects based on similarity or density of data objects. Ants are likely to pick up data objects that are either isolated or surrounded by dissimilar and then drop the picked ones in the similar types of data objects .Thus a proper clustering is achieved.

In this paper, we have analyzed the existing ant-clustering algorithm for diversified characteristics of the dataset. Numerical analysis method of linear equation is proposed based on the characteristics of the dataset that overcome the drawbacks of existing ant-clustering algorithm. The rest of the paper is laid out as follows. Section 2 states about the problem of existing ant-clustering algorithm. In section 3, related work and review of ant-clustering algorithm is briefly described. Our proposed method is presented in section 4. Section 5, explains the experimental results. Finally, Section 6 summarizes the study and concludes.

2. PROBLEM

There are motivations for using an ant-clustering algorithm in a clustering problem. In data clustering, many algorithms (like K-means and ISODATA) require an initial partition that is given as input before the data can be processed. But, "ant-clustering algorithm" does not require such an initial partitioning. "Ant-clustering algorithm" performs cluster analysis based on swarm intelligence. Swarm similarity is calculated for all the data objects, which measures the average similarity of each object with other objects. Then the swarm similarity is transformed to picking-up probability and dropping probability those are used to form the cluster. These picking-up probability and dropping probability uses two user defined parameters. This is one limitation of ant-clustering algorithm. It is hard to guess about the value of user defined parameters in advance to form the cluster because of the diversified characteristics of the dataset.

The main contribution of this paper is to calculate the picking-up probability and dropping probability based on the characteristics of the dataset that does not need to assign any user defined parameters in advance to process the data. Numerical analysis method of linear equation is proposed based on the characteristics of the dataset to form the cluster for diversified dataset.

3. RELATED WORK

There are two kinds of swarm intelligence methods used for clustering. One is ant colony optimization algorithm which is inspired by behaviors of ant colonies finding the shortest path between their nest and a food source [5-7]. The other is ant-based clustering inspired by the clustering of corpses and larval-sorting behaviors of real ant colonies [8-11].

In ant-clustering method, Ants are modeled by simple agents that randomly move in their environment, a 2D grid with periodic boundary conditions. Data objects that are scattered within this environment can be picked up, transported and dropped by the agents. The picking and dropping operations are based on the similarity and density of data objects within the ants' local neighborhood: ants are likely to pick up data objects that are either isolated or surrounded by dissimilar; they tend to drop them in the vicinity of similar ones. In this way, clusters of data objects on the grid are obtained.

For numeric data, a hybrid algorithm was proposed that discovers clusters automatically without prior knowledge of a possible number of classes, and without any initial partition [12]. It uses the stochastic and exploratory principles of an ant colony with the deterministic and heuristic principles of the K-means algorithm.

3.1. Review of Ant-Clustering Algorithm

The parameters and symbols used for ant clustering algorithm are illuminated as follows [2]:

α: swarm similarity coefficient;
r: observing radius of each ant;
N: the maximum of cycle times;
size: the size of the 2-dimension grid;
m_p: the number of ants in each population;
p: index of populations, p = 1, 2, 3;
Pp: picking-up probability;

p_d: dropping probability;

p_r: random probability, $p_r \in [0, 1)$;

k_1 and k_2: threshold constants for computing $_{\mathrm{Pp}}$ and p_d respectively;

ant_i: the ith ant;

o_i: the ith data object;

loaded and *unloaded*: state of ant. If there is a data object on an ant, its state is *loaded*; otherwise, its state is *unloaded*;

v_{high}: the speed of ants in high speed population;

v_{low}: the speed of ants in low speed population;

v_{MAX}: the maximal speed in variable speed population;

l: the maximum times of an ant moving with a same data object continuously.

3.2. Algorithm: A High-Level Description of Ant-Cluster

Initialization phase: Initialize parameters (α, r, N, *size*, m_p, v_{high}, v_{low}, v_{MAX}, and l). Place data objects on a 2-dimension grid randomly, i.e. assign a pair of coordinates (x, y) to each data objects. Put three populations of ants with different speed on this 2-dimension grid. Initial state of each ant is *unloaded*;

while (cycle_time <= N)

```
Adjust α with specific step;
for (p = 1; p <= 3; p++)
   for (i = 1; i <= m_p ; i++)
      if (ant_i encounter a data object)
         if (state of ant_i is unloaded)
            Compute the swarm similarity of the data object within a local region with radius
            r, and compute picking-up probability Pp. Compare Pp with a random probability
            p_r. if Pp > p_r, ant_i pick up this data object, and the state of ant_i is changed to loaded;
      else
         if (state of ant_i is loaded)
            If ant_i has already moved with the same data object l steps, the data object is
            dropped and the state of ant_i is changed to unloaded. Otherwise, compute the
            swarm similarity of the data object within a local region with radius r, and
            compute dropping probability p_d. Compare p_d with a random probability p_r. if p_d >
            p_r, ant_i drops this data object, and the state of ant_i is changed to unloaded.
   end
  end
end
```

3.3. General Description of Ant-Clustering Algorithm

In initialization phase, all parameters, including α, r, N, $size$, m_p, v_{high}, v_{low}, v_{MAX}, and l, are given values by user. Data objects and three populations of ants with different speed are placed in a 2-dimension grid randomly. There is only one data object and/or one ant in a grid at most. Initial state of each ant is set *unloaded*.

In each of outer loop iteration, i.e. while loop, all ants on the 2-dimension grid move one time. Each of interior loop iteration corresponds to the behavior of one ant. An ant moves one step on the 2-dimension grid randomly with different speed according to different population at a time. When it encounters a data object and its state is *unloaded*, the swarm similarity and picking-up probability are computed for deciding whether or not to pick up the date object. When it does not encounter a data object and its state is *loaded*, the moving times with the same data object is compared with l at first. If the ant has already moved with the same data object l times, the data object is dropped. Otherwise, the swarm similarity and dropping probability are computed for deciding whether or not to drop the date object.

The swarm similarity is computed by following formula:

$$f(0_i) = \frac{1}{S} \sum_{0_j \in Neigh(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha} \right] \qquad (1)$$

Where, 'o' represents each data object and $f(o_i)$ is a measure of the average similarity of object o_i with the other objects o_j present in the neighborhood of o_i. S is the number of objects o_j. $d(o_i, o_j)$ is the distance between two objects o_i and o_j in the space of attributes measured with Euclidean distance.

The swarm similarity is transformed to picking-up probability $_{Pp}$ and dropping probability p_d by following formulas respectively.

$$P_p(o_i)=\left(\frac{k_1}{k_1+f(o_i)}\right)^2,\ p_d(o_i)=\left(\frac{k_2}{k_2+f(o_i)}\right)^2 \quad (2)$$

Wherein, k_1 and k_2 are two user defined parameters. For diversified characteristics of the dataset, assigning value of k_1 and k_2 may not be able to form accurate cluster.

4. METHODOLOGY OF PRESENT WORK

The proposed calculating method of picking-up probability and dropping probability is carried out through the following steps:

Step 1: Calculate the distance of each object o_i with other objects o_j. $d(o_i, o_j)$ is the distance between two objects o_i and o_j in the 2D space of attributes measured with euclidean distance.

Step 2: Calculate the swarm similarity of each object by using equation 1.

Step 3: Find the relationship between data points and swarm similarity that came up with a linear relation. Express this relationship by linear equation.

Step 4. Calculate the error of swarm similarity for each object by fitting it in to the relation found from step 3 that is denoted by 'erswarm'.

Step 5. Calculate the mean of the error in swarm similarity.

$$meanerswarm = \sum erswarm / n$$

Step 6. Find the absolute difference of the error for each object from the mean error of similarity.

$$differ = absolute(erswarm - meanerswarm)$$

Step 7. Normalize the absolute difference from the mean difference of the error that is found from step 6.

$$dis\tan ce\_mean\_error = normalize\,(absolute\,(differ - mean\,(differ\,)))$$

Step 8. Sort the normalized data in an ascending order.

$$sorted\_dis\tan ce = sort\,(dis\tan ce\_mean\_error\,)$$

Step 9. Assign the picking-up probability of each data point such that each data point's picking-up probability will be its normalized absolute difference from the sorted distance for values [0 0.49].

Step 10. Assign the Dropping probability of each data point so that each data point's dropping probability will be its normalized absolute difference from the sorted distance for values[0.5 1].

$meanerswar\ m$, $differ$, $dis\tan ce\_mean\_error$, $sorted\_dis\tan ce$ are some intermediate variables to progress with our proposed method.

4.1. Description of proposed method

We have considered that, all the data points are scattered on 2D space. At first we have calculated the "euclidean distance" of each data point to all other data points in 2D space. For calculating the swarm similarity we have used the existing swarm similarity computation formula that was used in existing "ant-clustering algorithm". Then we tried to find out the relationship between swarm similarity and data points and we found that there is a liner relation between the swarm similarity

and distance of each data point to the other points on the 2D space. When we have plotted the swarm similarity and distance value of each object it came up with a liner relationship (for diversified characteristics of the dataset) which we can depicted as follows:

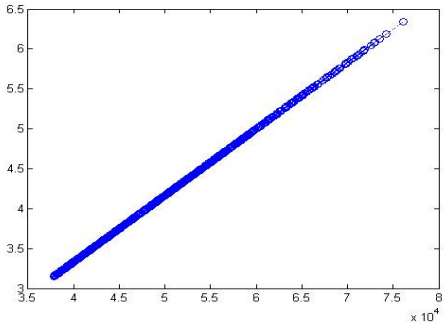


Figure 1. Swarm similarity and distance relationship for all data points

After finding the linear relationship between swarm similarity and distance of all data points we have calculated the error of swarm similarity for each object by fitting it in to the relation (that is represented by a linear equation). Then we have calculated the mean of the error in swarm similarity and found the absolute difference of the error for each object from the mean error of similarity. Then we have normalized the absolute difference from the mean difference of the error and sorted that in an ascending order. From this sorted data we have assigned the picking-up probability of each data point such that each data point's picking-up probability will be its normalized absolute difference from the sorted distance for values [0 0.49] and dropping probability for values [0.5 1].

In these ways, we do not need any user defined parameters to calculate the values of picking-up probability and dropping probability and we have got the values of picking-up probability and dropping probability within a range of 0 to 1 and always dropping probability was higher than picking-up probability. Our proposed method works for diversified datasets, because we have processed our method based on the characteristics of the datasets. Results of our proposed method are summarized in next section.

5. EXPERIMENTAL RESULTS

Recalling from the existing "Ant Clustering" algorithm [2], the two user defined values of K_1 and K_2 were respectively 0.10 and 0.15. They used larger value for K_2 to make the dropping probability larger to increase the effectiveness of the algorithm to form more accurate cluster. Then using these two user defined values, they gained P_p and P_d in the range of 0 to1. Finally these P_p and P_d are compared with P_r (random Probability) whose value (i.e. P_r) is also between 0 to1.Based on the value of P_p and P_d the actual cluster was formed.

In order to measure the performance of our proposed method we have used two synthetic datasets in 2D space. For all datasets, we have always got the linear relationship between swarm similarity and distance of all data points that looks like similar as fig. 1. We have coded our newly proposed method in MATLAB 7.

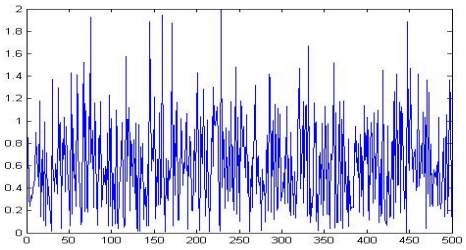
Dataset 1. This is a two dimensional data set having 500 points.
Dataset 2. This is a two dimensional data set having 1000 points.

In existing ant-clustering algorithm, it was suggested that swarm similarity coefficient (α) value must have to be between 12 to 14 to form the cluster. In our proposed method we have used different values of α for both datasets. The effects of different values of α on different datasets are listed in the followings tables.

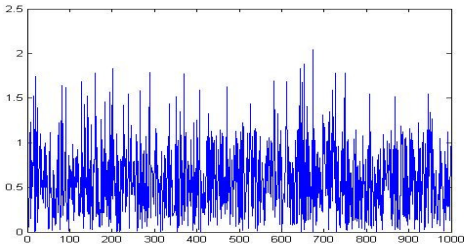
Table 1. Results of Synthetic Dataset (α=12)

No. of points	***No. of dissimilar data***	***meandiffer***	***Avg. of Pd***	***Avg. of Pp***
500	92	0.6062	0.9544	0.0356
1000	166	0.5811	0.9657	0.0242

In our proposed method, we have calculated error difference (denoted by 'differ') of each data point to the linear equation that is formed by distance and swarm similarity of all data points. Those data points have the 'differ' value more than 1 is considered as dissimilar data (outlier data). We have checked that, those data points have differ value more than 1 are positioned far away from the others data points. The 'differ' value of each data point is presented in the following figure:



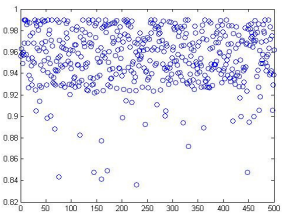
(a)



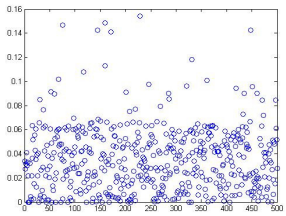
(b)

Figure 2. Error difference ('differ') value of synthetic datasets for α=12, (a) no. of data points 500, (b) no. of data points 1000.

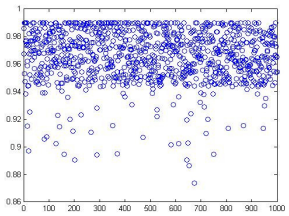
By using our proposed method we have got the value of picking-up probability $_{Pp}$ and dropping probability p_d within the range of 0 to 1. For both dataset 1 and dataset 2 our proposed method worked well.

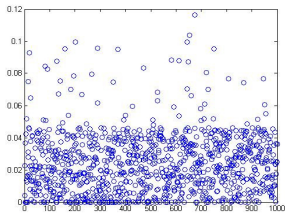


(a) p_d (500 data points)



(b) P_p (500 data points)



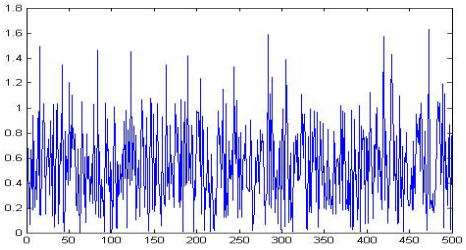


(c) p_d (1000 data points) (d) P_p (1000 data points)

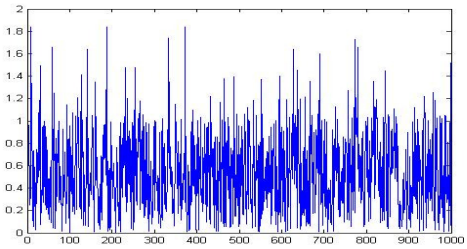
Figure 3. Dropping probability p_d and picking-up probability P_p values; (a) and (b) no. of data points 500, (c) and (d) no. of data points 1000.

Table 2. Results of Synthetic Dataset (α=13)

No. of points	***No. of dissimilar data***	***meandiffer***	***Avg. of Pd***	***Avg. of Pp***
500	49	0.5293	0.9537	0.0362
1000	101	0.5397	0.9649	0.0250

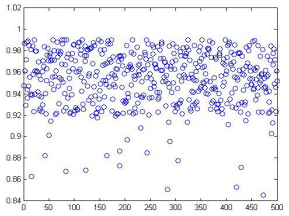


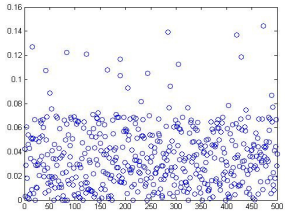
(a)



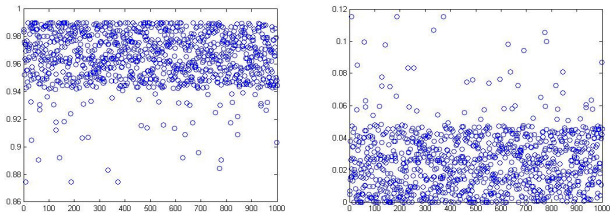
(b)

Figure 4. Error difference ('differ') value of synthetic datasets for α=13 (a) no. of data points 500, (b) no. of data points 1000.





(a) p_d (500 data points) (b) P_p (500 data points)

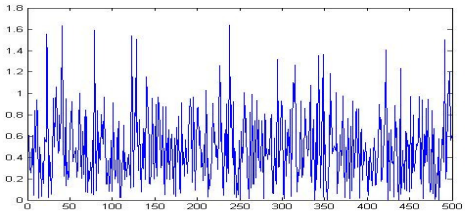


(c) p_d (1000 data points) (d) P_p (1000 data points)

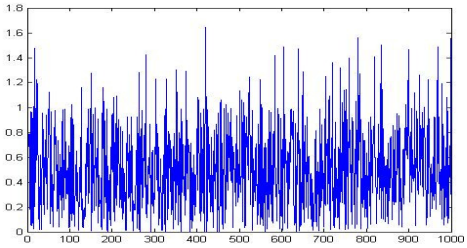
Figure 5. Dropping probability p_d and picking-up probability P_p values; (a) and (b) no. of data points 500, (c) and (d) no. of data points 1000.

Table 3. Results of Synthetic Dataset (α=14)

No. of points	***No. of dissimilar data***	***meandiffer***	***Avg. of Pd***	***Avg. of Pp***
500	26	0.4878	0.9550	0.0349
1000	62	0.5035	0.9652	0.0247

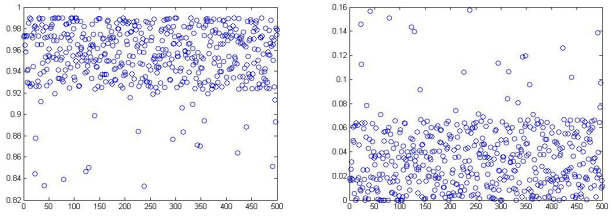


(a)

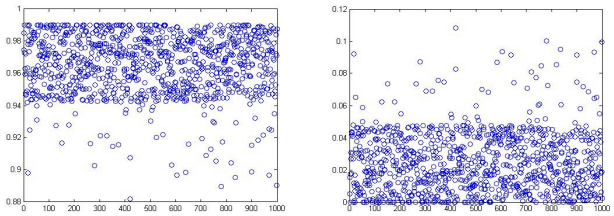


(b)

Figure 6. Error difference ('differ') value of synthetic datasets for α=14, (a) no. of data points 500, (b) no. of data points 1000.



(a) p_d (500 data points) (b) P_p (500 data points)



(c) p_d (1000 data points) (d) P_p (1000 data points)

Figure 7. Dropping probability p_d and picking-up probability P_p values; (a) and (b) no. of data points 500, (c) and (d) no. of data points 1000.

For both datasets (dataset 1 and dataset 2) and different values of swarm similarity coefficient (α= 12,13 and 14), our proposed method has the ability to find the values of picking-up probability and dropping probability within the range of 0 and 1. In existing "ant-clustering algorithm" they used larger value of K_2 to make the dropping probability larger to increase the effectiveness of the algorithm to get well formed cluster. By using our proposed method we have always got the larger dropping probability value than picking-up probability. (see table I, II, III and figure 3,5,7).

However, for larger value of swarm similarity coefficient (α = 14), "ant-clustering algorithm" has the ability to detect more similar data points. When we used smaller values of swarm similarity coefficient (α = 12 or 13), then number of dissimilar data points were larger. Larger number of dissimilar data points indicates that, it may not fall within any cluster boundary. If data points do not fall within any cluster boundary then we might not get well formed cluster. As we know that, clustering groups a set of similar data points to clusters. Therefore, reducing number of dissimilar data points increases the chance of falling more data points into the cluster to get well formed cluster.

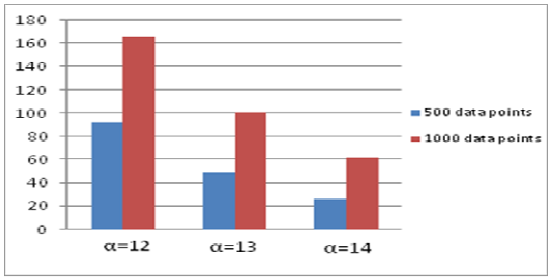


Figure 7. No. of dissimilar data points for different values of swarm similarity coefficient (α)

3. DISCUSSION AND CONCLUSION

Our proposed method uses the advantage of numerical analysis method of linear equation that overcomes the problem of using user defined parameters to calculate the picking-up probability and dropping probability to form the cluster. In our proposed method, we do not need any user defined parameters to calculate the values of picking-up probability and dropping probability and we have got the values of picking-up probability and dropping probability within a range of 0 to 1. It is important that, the dropping Probability be greater than the picking-up probability (P_d>P_p) so that an ant can drop the object (data) and get a new object (data) .Unless this constraint is preserved, we might not get well formed cluster. We have also observed that "ant-clustering algorithm" has the ability to detect more similar data for larger value of swarm similarity coefficients.

In each step of our calculation we also tried to see the effect of round-off error on our proposed method. But, we did not get any significant effect of round-off error to calculate the values of picking-up probability and dropping probability. Because, we have used only 500 and 1000 data points in our synthetic datasets. Therefore, it did not require more computations on which round-off error may have significant effect to form the cluster.

When data points are more scattered on 2D space it indicates diversified characteristics of the dataset. From our analytical result it is very evident that, our proposed method worked well for different characteristics of the dataset. Hence, in this paper, we dealt with only synthetic datasets and we believe our future work on real life dataset will also provide much accurate result as well.

ACKNOWLEDGEMENTS

Last but not the least, we would like to thanks all our friends and colleagues of our university. Also gratitude to IJDKP to give proper guidelines to format this paper.

REFERENCES

[1] J. Han and M. Kamber, (2001) *Data Mining Concepts and Techniques*, Morgan Kaufmann, San Francisco.

[2] P. jin, Y. zhu, and K. hu, (2007) "A clustering Algorithm for Data Mining Based on Swarm Intelligence," in *Proceedings of International Conference on Machine Learning Cybernetics.* (Hong Kong, 19-22 August).

[3] G.H. Ball and D.J. Hail, (1965) "ISODATA, a novel method of data analysis and pattern classification," Technical report, Stanford Research Institute.

[4] J. B. MacQueen, (1967) "Some Methods for classification and Analysis of Multivariate Observations," *Mathematical Statistics and Probability*, vol. 1, pp. 281-297.

[5] C. Tsai, C. Tsai, H. Wu, and T. Yang, (2004) "ACODF: a novel data clustering approach for data mining in large databases," Journal of Systems and Software, Vol. 73, pp. 133-145.

[6] C. Tsai, H. Wu, C. Tsai, (2002) "A new data clustering approach for data mining in large databases", in *Proceedings of the International Symposium on Parallel Architectures*, *Algorithms and Networks*. (Metro Manila, Philippines, pp. 22-24).

[7] H. Azzag, N. Monmarché, M. Slimane, C. Guinot, and G.Venturini, (2003) "AntTree: a new model for clustering with artificial ants," in *Proceedings of the 7th European Conference on Artificial Life.* (Dortmund, Germany, pp. 14-17).

[8] W. bin, S. Zhongzhi, (2001) "A clustering algorithm based on swarm intelligence," in *Proceedings of the International Conferences on Info-tech and Info-net*. (Beijing, China, pp. 58-66).

[9] W. Bin, Z. Yi, L. Shaohui, and S. Zhongzhi, (2002) "CSIM: a document clustering algorithm based on swarm intelligence," in *Proceedings of the Congress on Computational Intelligence*. (Hawaiian, USA, pp. 477-482).

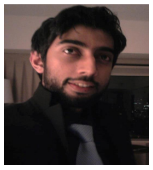
[10] Y. Yang, and M. Kamel, (2003) "Clustering ensemble using swarm intelligence," in *Proceedings of the IEEE Swarm Intelligence Symposium*Piscataway. (NJ, USA, pp. 65-71).

[11] J. Handl, J. Knowles, and M. Dorigo, (2003) "Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1D-som," Technical Report TR/IRIDIA/2003-24. IRIDIA, Universite Libre de Bruxelles, Belgium.

[12] N. Monmarché, M. Slimane, and G. Venturini, (1999) "On Improving Clustering in Numerical Databases With Artificial Ants," in Proceedings of the Conference on Advances in Artificial Life.

Authors Biography

Md. Asikur Rahman is a research based M.Sc. student at department of Computer Science at Memorial University of Newfoundland, Canada. He received his Bachelor's degree in B.Sc. (CIT) from Islamic University of Technology (IUT), Bangladesh, in 2008. His research interests are data mining, information visualization and visual analytics.



Md. Mustafizur Rahman is a research based M.Sc. student at Faculty of Engineering and Applied Sciences at Memorial University of Newfoundland, Canada. He received his Bachelor's degree in B.Sc. in Mechanical Engineering from Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2009. His research interests include Information Technology in Engineering, Numerical Simulation, and Advanced Control Systems.

Md. Mustafa Kamal Bhuiyan is a thesis based M.Sc. student at the department of Computer Science in Memorial University of Newfoundland, St. John's, NL, Canada. He received his Bachelor degree in Computer Science and Information Technology (CIT) from Islamic University of Technology (IUT), Bangladesh. His research interests are Information Visualization, Database Systems, Networking Architecture and Protocol.

S. M. Shahnewaz is a thesis based M.Sc. student at the department of Electrical and Computer Engineering (ECE) at University of Calgary, Canada. He has also completed his M.Sc. in Computer Science and Engineering from Islamic University of Technology (IUT). His research interests include automated software engineering, optimized software maintenance, data mining, and web technologies.