International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 02 | Feb -2017 www.irjet.net p-ISSN: 2395-0072

A Firefly based improved clustering algorithm

Priyanka Singhai, Prof Abhey Kothari, Mr. Rahul Moriwal

M.Tech, Computer Science &Engineering, Acropolis Institute of Technology & Research. Indore, M.P. India

----------****----------

Abstract—*The computational domain need to develop the methods by which the storage and data is handled effectively. Therefore the data mining techniques are utilized to evaluate the data and obtain the meaningful patterns to explore hidden knowledge. In this presented work the cluster data analysis technique is investigated. The cluster analysis is a technique by which the data is analysed in unsupervised manner to divide and decided the different groups of the data according to the user inputs. In this process the similarity among the grouped elements is the primary objective to achieve. This objective is help to find the better performance from the clustering algorithm.*

In this proposed work the clustering algorithm is studied in detail. Additionally the different clustering issues are addressed to achieve the good clustering. Finally the firefly optimization algorithm based clustering algorithm is followed for cluster data analysis. this technique is suffers from the long running time for performing the clustering therefore an improved clustering algorithm with the help of k-means algorithm and the firefly algorithm is proposed. The proposed technique provides ease in the centroid selection and the efficient and accurate data modeling. Additionally promises to reduce the processing time of the algorithm.

Further the proposed clustering technique is implemented with the help of visual studio environment. After implementation of the proposed algorithm the comparative study with the traditional firefly algorithm is performed. For comparative performance study the accuracy, error rate and resource consumption is taken as the primary parameters. The experimental results show the high performance outcomes during the data evaluation and accurate cluster formation.

Keywords—*data mining, cluster analysis, performance improvement, firefly algorithm, k-means.*

1. INTRODUCTION

The data mining is a domain of automatic data analysis. For evaluation of data there are two different approaches are used first supervised and second the unsupervised learning approach. In this presented work the supervised learning technique is used for investigation and demonstration. Data mining is a technique of analysing data and extraction of meaningful data for the real world applications. The extraction of data from the raw set of data needs to develop some computational data model by which the data is evaluated in certain criteria and return the matched data which is required by the application. The evaluation of data is performed in both the manners either with the supervisor or without the supervisor. In the machine learning and data mining the supervisor are the labelled data which is produced for analysis and using the class labels the learning process are keep in track. Most of the supervised learning algorithms are the classification algorithms and the unsupervised learning supports the clustering algorithms.

However the supervised learning algorithms are much accurate as compared to the unsupervised learning techniques. But the supervised learning techniques are always used with the labeled data and the amount of data is countable. On the other hand the unsupervised learning technique or clustering algorithms are used when the data is unlabelled or found in huge quantity. Therefore the proposed work is intended to explore the domain of data clustering and the performance improvement of the traditional clustering approaches.

Therefore the optimization based technique based technique namely firefly algorithm is used for investigation and solution design. Basically the clustering of data need to identify the optimal cluster centers using the optimization techniques. After finalizing the cluster centers the data clustering performed on the data. Therefore some initial improvement on the data centroid selection process is required to perform by which the solution becomes more effective and accurate for data analysis.

The data mining techniques are directly depends on the data which is used for analysis and pattern recovery. If the size of data is small, pre-defined classes are exist and the data is also refined and

cleaned then that is required to analyse such data using the classification algorithms which is a supervised learning approach. On the other hand if the data to be analysed is available in unstructured format, huge in quantity, available with some noisy contents then the supervised process is not suitable for data analysis. In this kind of data analysis the unsupervised learning technique or the data clustering is used for extracting the valuable patterns from the data.

In this presented work the main focus on the data mining based clustering algorithm is placed. The clustering algorithm is functioned on the data according to the similarity of the data elements and also the amount of clusters to be made. This process is not need to interact with the class labels to enhance the computed patterns. In study a number of clustering approaches are observed, but most of them either not much efficient in terms of processing time or ineffective for the accurate data analysis. In addition of that the noise in data can also affect the performance of the clustering algorithms such as the outliers or the missing attributes. Therefore a new technique is required to develop by which the efficiency and accuracy.

2. PROPOSED WORK

The proposed technique is based on the two step process of the cluster formation. Therefore first the data quality is enhanced and then the clustering approach is implemented on the refined data. During the pre-processing of the data the identical columns and missing attributes are also handled and then the well refined contents are processed for finding the optimum centroids of the data clusters. Finally the k-mean algorithm is used to allocate the suitable cluster data to the minimum distance centroids. The proposed algorithm is listed as follows:

Table 1 proposed algorithm

Enhanced firefly algorithm
Input: number of clusters K, input dataset D Output: K centroids, clustered data D_c Process: a. [Row, Col] = Read Data(D) // read the dataset and extract the dimensions of the data b. for $(i = 1; i \leq Row; i++)$ // elevate the dataset for all the row and columns which contains the null values or missing values and remove it 1. for $(j = 1; j \leq Col; j++)$ i. if $D(i,j) == null$ 1. $remove(D(i).row)$ ii. end if 2. end for c. end for d. n 1. $if\ unique(D(k)) == 1$ i. Remove $(D(k))$ 2. End if 3. $if\ unique(D(k)) == Row$ i. Remove $(D(k))$ 4. End if e. End for f. Initialize the firefly $FF_{init}[]$ g. [R, C] = SizeOf($FF_{init}[\ ]$) h. While number of iterations 1. $for(i = 1; i \leq R; i++)$ 2. $C_{data} = FF_{init}[i]$ //selecting firefly i. $for(j = 1; j \leq Row; j++)$ 1. $D_{row} = D[j]$ 2. $Fitness[C_{data}, i] = \sqrt{\sum_{j=1}^{N}(C_{data} - D_{row})^2}$ ii. End for 3. End for 4. Select best points from $Fitness[C_{data}, i]$ 5. $if(best\ points == k)$ i. $return\ while$ 6. End if 7. Go to step 8 i. End while j. Allocate the centers to the datasets

3. RESULT ANALYSIS

The chapter provides the evaluated results and the comparative study among new technique proposed and traditional firefly algorithm. This chapter helps to understand how the proposed approach performing better than the traditional technique.

3.1 Accuracy

The accuracy of the algorithm provides the estimation about accurately distinguishing the groups of data. Therefore that is an essential parameter for any data analysis algorithm. This parameter can be evaluated using the following formula.

$$accuracy = \frac{correctly\ identified\ samples}{total\ samples} X100$$

Table 2 Accuracy

Dataset size	Proposed algorithm	Firefly algorithm
50	77.93	55.29
100	81.37	58.34
200	82.58	59.17
300	85.32	60.63
500	86.16	62.56
700	87.63	64.31
1000	89.92	65.42

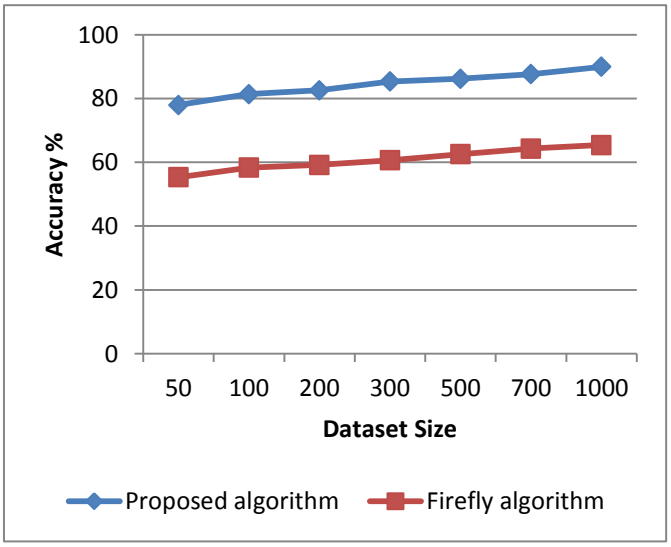


Chart- 1 Accuracy

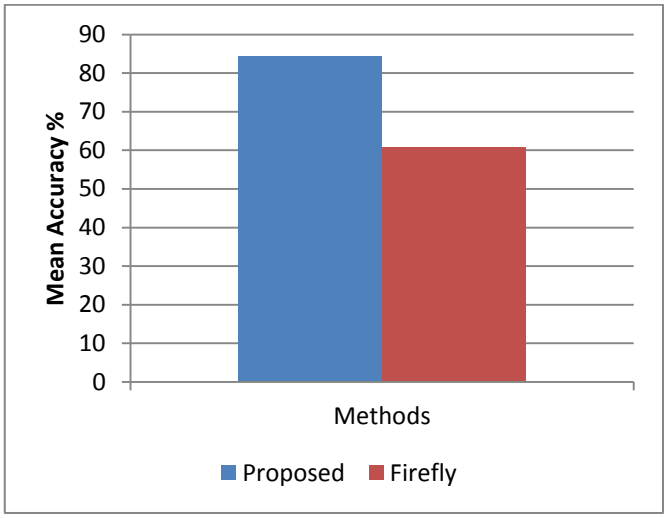


Chart- 2 mean accuracy

Chart 1 and the table 2 shows the evaluated performance in terms of accuracy. In this figure the amount of dataset instances for evaluation is given in X axis and the Y axis shows the percentage accuracy obtained by the system. According to the obtained performance the accuracy of the proposed clustering algorithm is efficient as compared to the traditional firefly algorithm. Additionally produces constant accuracy as compared to the traditional method. In order to justify the results more clearer the mean accuracy of both the algorithms are evaluated and demonstrated in figure 2. According to this diagram the X axis contains the methods implemented and the Y axis shows the mean accuracy percentage. The combined results over the different size of dataset shows the higher percentage of gain as compared to the traditional method additionally able to produce the accuracy 78-89%.

3.2 Error rate

The error rate is an amount of data that is not properly recognized during the automated data analysis. That can be evaluated using the following formula:

$$\text{error rate} = 100 - \text{accuracy}$$

Or

$$\text{error rate\%} = \frac{\text{total misidentified data}}{\text{total samples produced}} \text{X100}$$

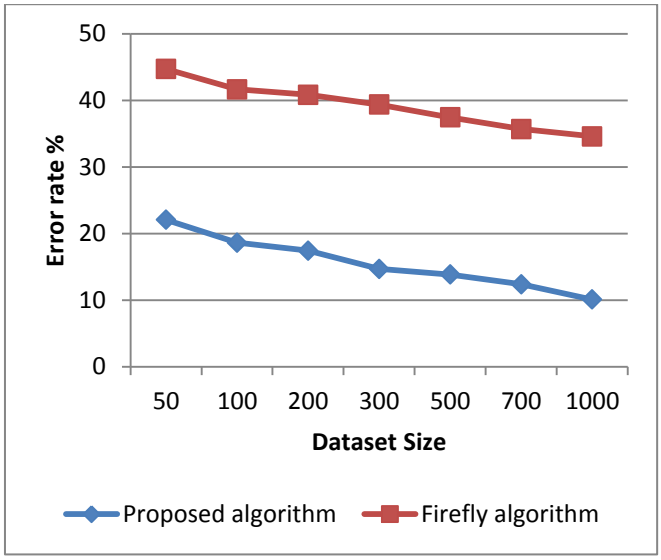


Chart- 3 Error rate

Table 3 Error rate

Dataset size	Proposed algorithm	Firefly algorithm
50	22.07	44.71
100	18.63	41.66
200	17.42	40.83
300	14.68	39.37
500	13.84	37.44
700	12.37	35.69
1000	10.08	34.58

The error rate in terms of percentage of both the implemented algorithms is given using table 3 and figure 3. In this figure the size of datasets is given using X axis and the Y axis shows the percentage error rate. According to the given results the proposed technique produces less error rate as compared to the traditional technique of clustering optimization. Additionally the error rates of both the systems are reducing that is a good significant with increasing size of data. In addition of that for justifying the results the mean error rate percentage is also estimated. The mean error rate percentage is given using the figure 4. In this diagram the X axis shows the implemented techniques and the Y axis shows the mean error rate percentage. According to the obtained results the proposed technique produces less error rate as compared to the traditional algorithm. Thus the proposed technique is much adoptable than classical approach.

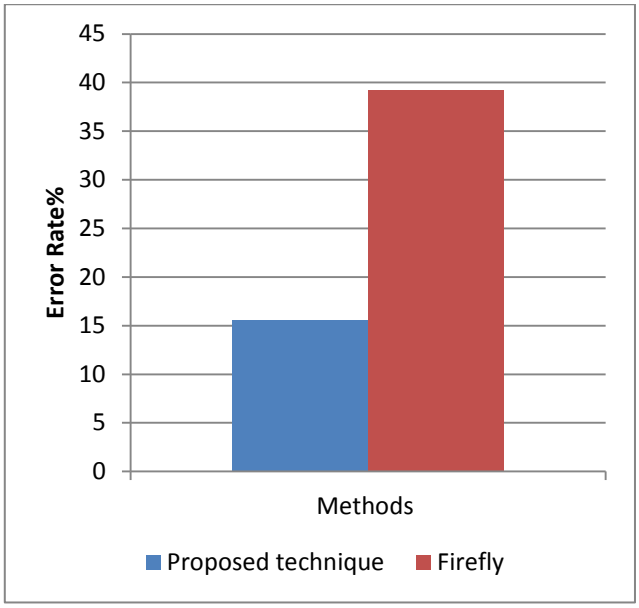


Chart- 4 mean error rate

3.3 Memory usage

The amount of main memory required to evaluate the data using the given algorithm is known as the memory usage of the algorithm. The figure 5 and table 4 shows the memory

Table 4 Memory usage

Dataset size	Proposed algorithm	Firefly algorithm
50	27817	26801
100	28865	26615
200	29629	27844
300	30562	29217
500	31983	30174
700	32717	31831
1000	33104	32947

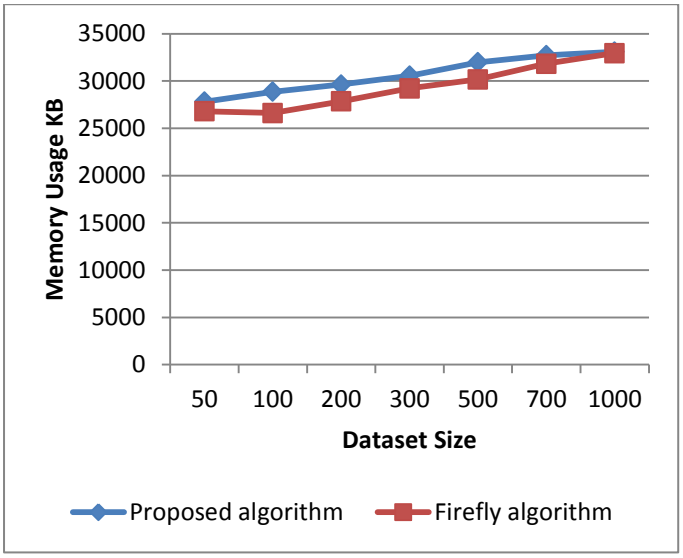


Chart- 5 Memory usage

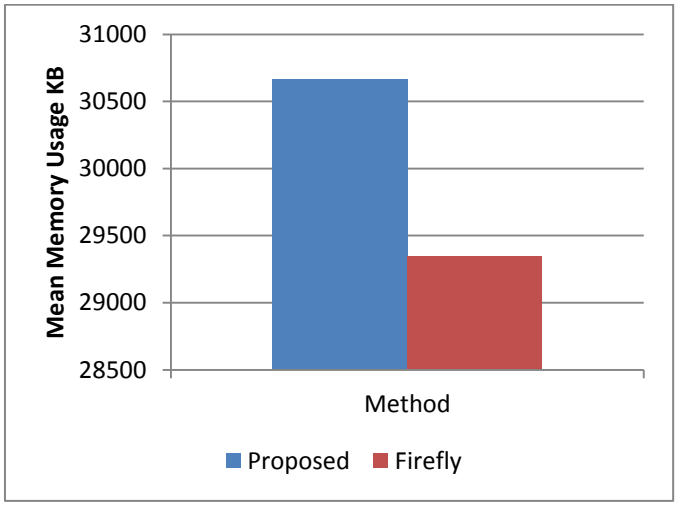


Chart 6-mean memory usage

Usage of the implemented algorithms, the according to the given figure 5 X axis contains the amount of data to be process in increasing size. Similarly the Y axis contains the memory usage in terms of KB. According to the given results the memory requirement of the proposed system is higher as compared to the traditional firefly optimization technique. To understand the difference among both the technique's memory requirements figure 6 shows the mean memory consumption of both the algorithms. In this diagram the X axis contains the implemented methods and the Y axis shows the mean memory consumption of the algorithms. The mean results show the performance of the traditional algorithm is much effective as compared to the traditional algorithm.

3.4 Time consumption

The amount of time required to evaluate the entire data and produces the clusters are given here as the time consumption of the algorithm. The utilized time of the algorithms are evaluated and given in terms of milliseconds. In this diagram the Y axis contains required time in MS and the X axis contains the amount of data to be processed. According to the given results the proposed method consumes less amount of time as compared to

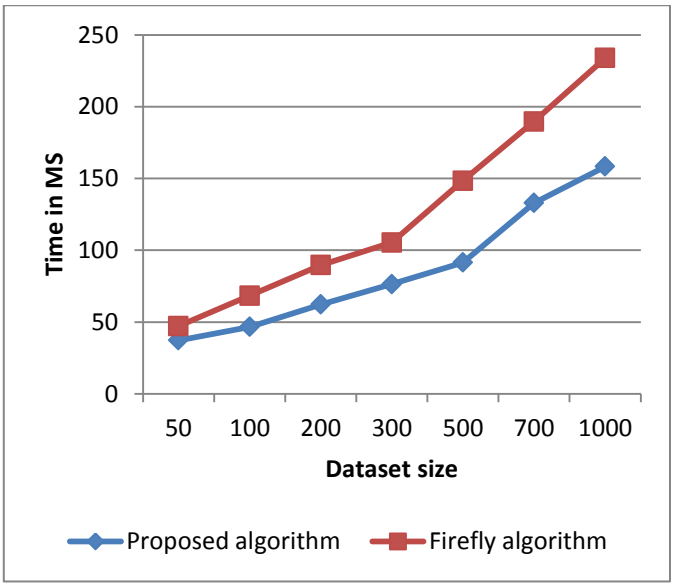
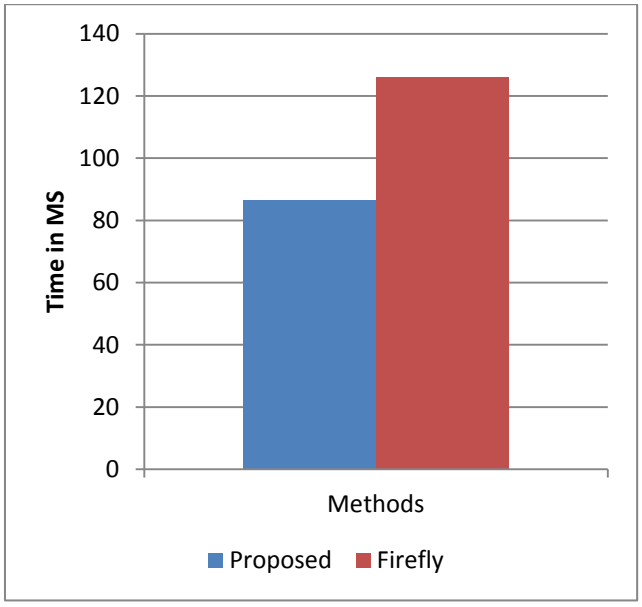


Chart-7 Time consumption

Table 5 Time consumption

Dataset size	Proposed algorithm	Firefly algorithm
50	37.27	47.22
100	46.59	68.36
200	62.18	89.61
300	76.41	105.32
500	91.47	148.38
700	132.92	189.66
1000	158.34	233.95



Charrt 8- mean time consumption

Traditional firefly optimization based clustering approach. In addition of that for clarifying the results obtained the mean performance of both the techniques are reported using the figure 8. In this diagram the X axis shows the implemented algorithms and the Y axis shows the mean time consumption of the given algorithm. According to the mean performance of the algorithms the proposed technique consumes less time as compared to the traditional approach of firefly algorithm.

4. CONCLUSION

The proposed work is indented to find an efficient clustering scheme for enhancing the clustering accuracy and optimization time. The primary goal of the clustering scheme development is achieved and this chapter provides the entire summary of the conducted study work. In the addition of that the future extension of the work is also involved after conclusion of the work.

4.1 Conclusion

The automated data analysis techniques are becomes more crucial now in these days, because a huge amount of records in unstructured format is generated every day. For automatic data analysis data mining provides different algorithms and tools that are help to evaluate the data and classify them or cluster them. Among the different approaches when the data is found in large quantity and also the nature of data is unlabelled then the supervised learning approaches are not suitable for use. Therefore the clustering of the data is a good strategy for performing the data analysis. In this presented work the data mining based clustering algorithm is studied in detail. In addition of that a new clustering algorithm is also proposed for enhancing the centroid selection of the clustering.

The proposed technique is a technique where first the data is pre-evaluated and pre-processed for improving the data quality. After that the outlier points are recovered and removed from the input data. After enhancing the quality of data the data normalization is performed to scale entire data in a similar scale and finally the optimal cluster centers (centroids) are estimated. Finally these centroids are used to perform clustering in data or making groups of the data point available. The given technique usages the traditional firefly algorithm for estimating the actual optimum cluster centers and the Euclidean distance is used to find the other cluster data points that are belongs to the obtain centroids. This approach promises to provide the accurate clustering of data by improving the data quality and optimum centroid selection.

The implementation of the proposed clustering technique is performed using the visual studio technology and their performance in terms of accuracy, error rate, memory consumption and the time consumption is estimated. According to the obtained results the proposed technique is efficient and accurate as compared to the traditional firefly based clustering approach. The obtained performance is summarized using the given table 6.

Table 6 Performance summary

S. No.	Parameters	Proposed technique	Firefly algorithm
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory consumption	High	Low
4	Time complexity	Low	High

According to the obtained performance the proposed algorithm is accurate and efficient as compared to the traditional firefly based clustering algorithm.

Therefore the proposed technique is adoptable as compared to the traditional technique.

4.2 Future work

The proposed work is an enhanced algorithm for performing clustering based on firefly optimization technique and traditional k-mean clustering algorithm. The proposed technique enhances the algorithms clustering accuracy and time consumption but consumes additional main memory during the data evaluation. Therefore need to enhance the algorithm for memory consumption. Additionally the given algorithm is not utilized with the real world application yet. Thus need to test before use with a real world application such as image segmentation and other large scale data. x.

REFERENCES

[1] Tahereh Hassanzadeh, Mohammad Reza Meybodi, "A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means", 2012 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)

[2] Data Mining: What is Data Mining?, http://www.anderson.ucla.edu/faculty/jason. frand/teacher/technologies/palace/datamining.htm

[3] Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[4] MahakChowdhary, ShrutikaSuri and MansiBhutani, "Comparative Study of Intrusion Detection System", 2014, IJCSE All Rights Reserved, Volume-2, Issue-4

[5] Mrs. PradnyaMuley, Dr. Anniruddha Joshi, "Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 4, Volume 2 (April 2015)

[6] GhazalehKhodabandelou, Charlotte Hug, Rebecca Deneckere, Camille Salinesi, "Supervised vs. Unsupervised Learning for Intentional Process Model Discovery", Business Process Modeling, Development, and Support (BPMDS), Jun 2014, Thessalonique, Greece. pp.1-15, 2014

[7] K. Jayavani, "STATISTICAL CLASSIFICATION IN MACHINE INTELLEGENT", ISR Journals and Publications, Volume: 1 Issue: 1 18-Jul-2014, I

[8] JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[9] Chih-Feng Chao and Ming-HuwiHorng, "The Construction of Support Vector Machine Classifier Using the Firefly Algorithm", Hindawi Publishing Corporation Computational Intelligence and Neuroscience Volume 2015, Article ID 212719, 8 pages

[10] Chih-Feng Chao, Ming-HuwiHorng, and Yu-Chan Chen, "Motion Estimation Using the Firefly Algorithm in Ultrasonic Image Sequence of Soft Tissue", Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2015, Article ID 343217, 8 pages

[11] Jianjun Zhang, Yueguang Li, "An Improved Firefly Algorithm and its application in Time-table Problems", International Symposium on Computers & Informatics (ISCI 2015)

[12] Minmei Huang, Jijun Yuan, and Jing Xiao, "An Adapted Firefly Algorithm for Product Development Project Scheduling with Fuzzy Activity Duration", Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2015, Article ID 973291, 11 pages

[13] L.E. Agustı´n-Blas, S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, J.A. Portilla-Figueras, "A new grouping genetic algorithm for clustering problems", 2012 Elsevier Ltd. All rights reserved.

[14] AbdolrezaHatamlou, "Black hole: A new heuristic optimization approach for data clustering", 2012 Elsevier Inc. All rights reserved.

[15] Tang Rui, Simon Fong, Xin-She Yang, Suash Deb, "Nature-inspired Clustering Algorithms for Web Intelligence Data", 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

[16] O.A. Mohamed Jafar, R. Sivakumar, "A Study of Bio-inspired Algorithm to Data Clustering using Different Distance Measures", International Journal of Computer Applications (0975 – 8887) Volume 66– No.12, March 2013

[17] Rui Wang, Robin C. Purshouse, Peter J. Fleming, "Local Preference-inspired Co-evolutionary Algorithms", GECCO'12, July 7–11, 2012,

Philadelphia, Pennsylvania, USA Copyright 2012 ACM 978-1-4503-1177-9/12/07

[18] DervisKaraboga, CelalOzturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", © 2009 Elsevier B.V. All rights reserved.

[19] An improved K-Means clustering algorithm, Juntao Wang, Xiaolong Su, 978-1-61284-486-2/111$26.00 ©2011 IEEE