International Journal of Soft Computing 7 (2): 71-78, 2012
ISSN: 1816-9503

A Cluster-Based Deviation Detection Task Using the Artificial Bee Colony (ABC) Algorithm

M. Faiza Abdulsalam and Azuraliza Abu Bakar
Center for Artificial Intelligence Technology, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Abstract: The Artificial Bee Colony (ABC) algorithm was motivated by the intelligent foraging behavior of honey bee swarms. The ABC algorithm was developed to solve clustering problems and revealed promising results in processing time and solution quality although, no research has yet considered employing ABC for deviation detection. In this study, researchers propose modifying the ABC clustering algorithm for deviation detection. An outlier factor has been used to identify the top n outliers that deviate from the dataset. The proposed algorithm was tested on three UCI benchmark datasets. Experimental results have shown that the ABC deviation detection algorithm has performed with comparable results.

Key words: Clustering, deviation detection, clustering-based deviation detection, Artificial Bee Colony (ABC) algorithm, intelligent, Malaysia

INTRODUCTION

Deviation detection is a primary step in many data mining applications. It includes determining a set of observations whose values deviate from the expected range. These extreme values could improperly impact the analysis results and thus lead to incorrect results (Tiwari *et al*., 2007). Detecting and removing deviations are important data mining tasks. Errors in large databases can be extremely common so an important property of a data mining algorithm is robustness with respect to deviations in the database (Mansur and Sap, 2005). Deviation detection aims to find the infrequent data with exceptional behaviors compared with other data. It has been researched within numerous application domains and knowledge disciplines and has become increasingly beneficial in some applications including credit card fraud detection, calling card fraud detection, intrusion detection and discovery of criminal activity (Jiang and Yang, 2009).

Most developed data mining methods address this problem to some extent but not fully and they can be improved by addressing the problem more directly. Deviation detection can be regarded as a task complementary to clustering. There are many approaches to detect deviations including statistical-model, clustering, distance and density-based approaches. This study considers the clustering-based approach for detecting deviations. Clustering-based approaches aim to partition data into a number of clusters where each data point can be assigned a membership degree for each cluster. For deviation detection, they consider deviation so, it does not interfere with the clustering process (Mansur and Sap, 2005). Some existing algorithms including DBSCAN (Ester *et al*., 1996), CLARANS (Ng and Han, 1994) and BIRCH (Zhang *et al*., 1996), consider deviations similarly. However, others algorithms have been adapted to identify the deviations through post-processing process. Some researchers have extended K-means and K-medoids. Conversely, others used wavelet transforms to detect deviations by progressively removing clusters from the dataset (Ceglar *et al*., 2007).

Cluster analysis as a data mining task is an active research topic in many fields including machine learning, statistics, data mining and pattern recognition. It aims to group data points into a set of clusters (or groups) such that data in the same cluster share a high degree of similarity while being dissimilar to data in other clusters (Fathian *et al*., 2007). Generally, clustering algorithms can be divided into two categories: partitional and hierarchical clustering. Hierarchical clustering is not the interest in this study which instead focuses on partitional clustering. Partitional clustering algorithms divide data into a predefined number of clusters by optimizing certain criteria (Zou *et al*., 2010). One of the most common partitional clustering algorithms is the K-means algorithm which is a center-based clustering algorithm. During the

Corresponding Author: M. Faiza Abdulsalam, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

last three decades, the clustering K-means algorithm has been used because of its simplicity and high speed in clustering large datasets. However, the K-means algorithm has two drawbacks: it converges to the local optimum solution and is sensitive to the initial states (Selim and Ismail, 1984). To address this problem, researchers have proposed many population-based stochastic search algorithms such as the Ant Colony Optimization (ACO) and Artificial Bee Colony (ABC) algorithms.

The ABC algorithm is one of the most recently presented swarm-based algorithms. Karaboga (2005) first presented the ABC algorithm at the Erciyes University of Turkey for numerical optimization problems based on the foraging behavior of honey bee swarms. The ABC algorithm has been implemented to solve many problems. It was applied to solve clustering problems which shows that clustering and deviation detection are closely related tasks. This research aims to propose a new Clustering-Based Deviation Detection Method using ABC algorithm. This method has great potential as a new method in detecting deviations after getting clustered data from large and high dimensional datasets. This new method is also expected to be efficient and able to detect deviations correctly at good speeds. This study uses two benchmark measurements to detect deviations. These measurements are the Detection Rate (DR) and False alarm Rate (FR). Two Cluster-based Deviation Detection methods are used for comparison with and evaluation of the proposed method.

MATERIALS AND METHODS

Related works: Clustering and deviation detection are closely related tasks. From the clustering perspective, deviations are objects that do not belong to any cluster in deviation detection these objects may appear as deviations. Moreover, from the deviation detection perspective, clustering has drawn attention from many researchers. The clustering process becomes more difficult however when the data include points that do not belong to any cluster. Many researchers have argued whether clustering algorithms are an appropriate choice for the deviation detection process. Zhang and Wang (2006) stated that clustering algorithms should not be considered as Deviation Detection Methods. This might be true for some clustering algorithms such as the K-means clustering algorithm because the cluster means that it produces are sensitive to noise and deviations (Laan *et al*., 2003).

Literature on the deviation detection problem often describes a variety of approaches and techniques that attempt to solve this problem. The clustering-based approach is one of the most commonly studied approaches for deviation detection. The main concern of clustering-based deviation detection algorithms is to find clusters and deviations which are often regarded as noise that should be detected and removed to make the clustering process more reliable (He *et al*., 2003).

Niu proposed a distance-based deviation definition and detection algorithm DCOD (Distribution Clustering Outlier Detection). The researchers redefined the problem by clustering in the distribution difference space instead of the original feature space. Consequently, the new algorithm is stable despite different input and scalable to dimensionality. Experiments on both real and synthetic datasets showed that DCOD outperformed its counterpart in both efficiency and effectiveness. Jiang *et al*. (2010) extended the idea about an object's outlier factor to the cluster situation. A clustering-based deviation detection approach, CBOD is introduced by Jiang and An (2008) according to a cluster outlier factor. The method includes two phases: the first phase clusters the dataset using a one-pass clustering algorithm and the second phase determines outlier clusters by calculating the outlier factors.

For high-dimensional data however, objects may belong to different clusters in different subspaces. More fine-grained concepts to define deviations are therefore needed. Seidl *et al*. (2009) addressed deviation detection in heterogeneous high-dimensional data and proposed a new OutRank approach using a novel scoring function that provides a consistent model for ranking deviations in the presence of different attribute types. Su described a deviation detection algorithm by showing the arbitrary shape clustering approach and explaining the abnormal cluster notion. The algorithm first divides the dataset into many clusters using the suggested clustering approach. Deviations are then discovered from the cluster set based on the abnormal cluster concept. By applying inter-cluster dissimilarity measurement, the proposed algorithm has an acceptable performance on the mixed data.

Al-Zoubi *et al*. (2010) presented a method based on fuzzy clustering approaches for deviation detection. First, a c-means fuzzy clustering algorithm is performed. Small clusters are then determined and considered as deviation clusters. The remaining deviations (if any) are then detected in the remaining clusters based on temporarily removing an object from the data set and re-calculating the objective function. If a noticeable change occurs in the Objective Function (OF), the object is considered a deviation.

Over the last decade, research has been directed towards modeling the behavior of some social insects such as ants and bees for many purposes including

problem searching and solving. Meanwhile, population-based optimization algorithms have been implemented to solve the clustering problem. Using an ant colony is an ideal swarm-based optimization approach where the search process is inspired by modeling real ants behaviors. Some approaches have been proposed to model the intelligent behavior of honey bees and applied to solve the clustering problem (Zhang *et al*., 2010). Alam *et al*. (2010) proposed using a novel swarm intelligence-based clustering technique for deviation detection, called Hierarchical Particle Swarm Optimization-Based Clustering (HPSO-clustering). The proposed technique can perform Hierarchical Agglomerative Clustering (HAC) and deviation detection. In the proposed approach, a swarm of particles evolves through different stages to identify deviations and normal clusters.

Mohemmed *et al*. (2010) also applied Particle Swarm Optimization (PSO) to solve the deviation detection problem. The new PSO approach is compared with a generally used Deviation Detection Method, the Local Outlier Factor (LOF) and the experiments were performed on five real data sets. The results showed that the novel PSO Method extremely outperformed the LOF approach in correctly identifying the deviations on most datasets. This study extends the Artificial Bee Colony (ABC) scope and offers a new insight into the deviation detection problem. The deviation detection problem is converted into an optimization problem. This study employs the ABC algorithm to solve the deviation detection problem.

Clustered-based deviation detection using the ABC algorithm: Many approaches have been used to detect deviations the clustering-based approach is one of the most commonly developed approaches because clustering is a popular technique used in deviation detection. This study uses the Clustering-Based Deviation Detection Method and the Artificial Bee Colony (ABC) algorithm is applied to detect clustered-based deviations. The proposed method is performed using two consecutive stages that are repeated several times. In the first stage, the dataset is clustered into a set of clusters, C_i, i = 1, 2, 3, ... , k and the deviated objects are detected in the second stage by assigning an outlyingness factor to each object. The associated factor depends on the object's distance from the centroid of the cluster.

The first stage involves deploying the ABC algorithm to cluster the dataset into a specific number of clusters k. In the ABC algorithm, the position of a food source is represented as a possible solution to the problem to be optimized and the amount of nectar retrieved from a food source conforms to the q uality of the solution represented by that food source. Similar to other swarm intelligence-based algorithms, the ABC algorithm works iteratively. Assuming the number of food sources is NS which equals the number of employed or onlooker bees and D is the dimension of each solution vector (Karaboga and Akay, 2009; Karaboga and Basturk, 2008). The main steps of the ABC algorithm can be explained:

Step 1: The first step is initialization. In this step, a set of food source positions are generated and distributed randomly. This means initializing a random population of NS solutions (S_1, S_2, ..., S_{NS}) where $S_i = \{s_{i1}, s_{i2}, ...,s_{id}\}$; each solution S_i is a D-dimensional vector. The maximum and minimum values of each vector are determined to calculate the initial solutions (initial cluster centroids) using the following Equation:

$$S_{ij} = x_{min_j} + \text{rand}\,[0, 1] \times (x_{max_j} - x_{min_j}) \quad (1)$$

where, x_{min_j} and x_{max_j} represent the lower and upper bounds of dimension j, respectively. After initializing the solution population, the fitness value of each food source position (solution) is evaluated. In the ABC algorithm, the fitness of each solution is the value of the objective function (the objective clustering function). The ABC algorithm has some control parameters: NS, the colony size (employed bees and onlooker bees); Food number (the number of food sources equals half of the colony size); limit; MR, the Modification Rate and MCN, the Maximum Cycle Number. Initially, the values of these control parameters are assigned.

Step 2: Each employed bee searches the neighborhood of its current food source to determine a candidate food source position:

$$Z_{ij} = \begin{cases} x_{ij} + \theta\,(x_{ij} - x_{jk}) & \text{if } r_j < MR \\ x_{ij} & \text{otherwise} \end{cases} \quad (2)$$

where, $j \in \{1, 2, \ldots, D\}$ and $k \in \{1, 2, \ldots, NS\}$ are randomly chosen indexes. k must differ from i. θ is a random number between [-1, 1]. MR is a Modification Rate this parameter is a control parameter that controls whether element x_{ij} will be modified and r_j is randomly chosen number in the range [0, 1].

Step 3: After generating the new food source, its nectar amount will be evaluated and a greedy selection will be performed. If the quality of the newly generated food source position is better than the current food source position, the employed bee leaves the old food source

position and moves to the new one. If the fitness of the new food source is equal to or better than that of S_i, the new food source takes the place of S_i in the population and becomes a new member.

Step 4: An onlooker bee first selects a food source by evaluating the information received from all employed bees. The probability p_i of selecting food source i is determined using the following expression:

$$p_i = \frac{f_i}{\sum_{i=1}^{NS} f_i} \quad (3)$$

where, f_i represents the fitness value of food source S_i. After selecting a food source, the onlooker bee generates a new food source using Eq. 4. Once the new food source is generated, it will be evaluated and a greedy selection will be applied as for employed bees.

Step 5: If a predetermined number of trials cannot further improve a candidate solution represented by a food source then that food source is considered abandoned and the employed bee associated with that food source becomes a scout. The scout bee randomly generates a new food source using Eq. 4:

$$v_{ij} = x_{min_j} + rand\,[0,1] \times (x_{max_j} - x_{min_j}) \text{ for } j = 1, 2, 3, \ldots, D \quad (4)$$

The abandoned food source is replaced by the randomly generated new food source. In the ABC algorithm, the predetermined number of trials for abandoning a food source is called the limit. This is considered one of the control parameters. No more than one employed bee can become a scout in each cycle in this algorithm.

Step 6: If a termination condition is reached, the process is stopped and the best food source position is returned otherwise the algorithm returns to Step 2 (Karaboga and Akay, 2009; Karaboga and Basturk, 2008).

A set of k cluster centers specifies the object partitions by mapping the cluster search space to the partition search space. The partition is identified by assigning each object to the cluster that is associated with its nearest cluster center. The nearest refers to a distance metric which is the Euclidean distance used in this study (Fig. 1 shows the distance between each object and its closest cluster center).

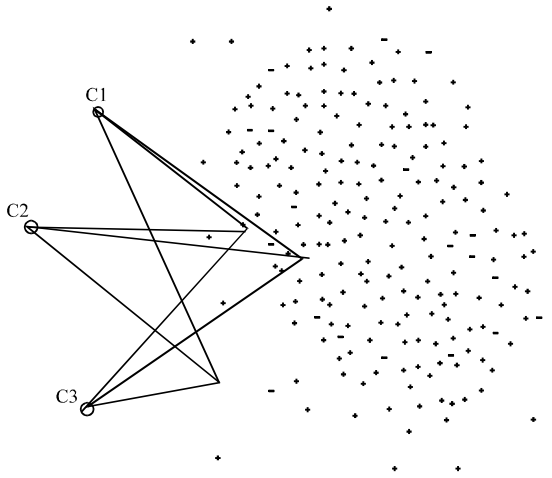

Fig. 1: The distances between data points and cluster centroids

In the clustering process, each solution represents the cluster centroids and the distance between each data point and the cluster center is calculated using Euclidean distance measures. Specifically, each data point is assigned to the nearest cluster (cluster centroid). To evaluate the clustering validity, intra-cluster similarity is calculated for every cluster.

The second stage detects outlying objects in the dataset by calculating the outliers factor values for every object. This factor depends on the distance from the object to the centroid of the cluster to which the object belongs. The algorithm starts iteratively by first finding the object with the maximum distance d_{max} to the cluster centroid thus:

$$d_{max} = max_i \{\| x_i - C_i \|\}, i = 1, 2, 3, \ldots .1\ N \quad (5)$$

Outlier factors o_i, for every object are then calculated. An outlier factor value for object x_i is calculated using this equation:

$$o_i = \frac{\| x_i - C_i \|}{d_{max}} \quad (6)$$

Where:

$\|x_i - C_i\|$ = The distance between every object x_i and its allocated cluster centroid C_i

d_{max} = The maximum distance of a certain object to the cluster center [7]

After all iterations, every object will have an outlier factor value that represents the object's deviation degree. All outlier factor values of the dataset are normalized to the range [0, 1]. The outlier factor value is compared with a predefined threshold value T that lies between 0 and 1. An outlier factor with a greater value is more likely to be a deviation. The object for which o_i>T is considered a

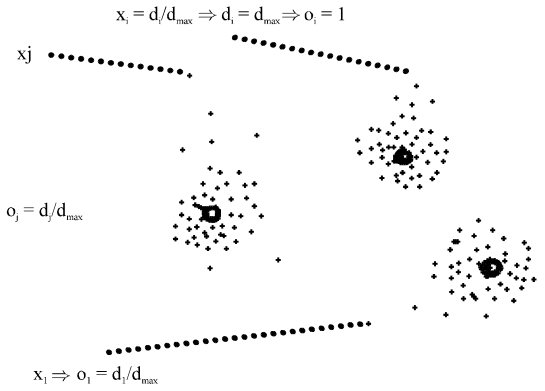

Fig. 2: Example of outlier factors

deviation. For demonstration, Fig. 2 shows an example that shows a dataset clustered into three clusters with their calculated outlier factors.

RESULTS AND DISCUSSION

The evaluation of deviation detection approaches includes particular difficulty as a deviation has distinct definitions and different discipline experts regularly have individual perspectives on whether a detected case is a true deviation. This study considers unbalanced datasets where some or all objects from the minority class are selected as deviations while objects from all or some instances of the majority classes comprise the normal set. In some datasets, some instances of the majority classes are so, distinct that they can be detected as deviations. A comprehensive performance study has been conducted to evaluate the proposed method. The datasets used in the experiments are Wisconsin Breast Cancer (WBC), Glass (GLS) and Lymphography (LYMP) datasets. These datasets were obtained from the UCI Machine Learning Repository (Blake and Merz, 2010). The deviations have been identified according to the outlier factor value of each object because the datasets are different and each has its own characteristics (number of instances, number of attributes). Different values between 0 and 1 have thus been assigned to the outlier factors. Two measurements have been used to measure the performance of the proposed algorithm: Detection Rate (DR) and False Alarm Rate (FR). The results of every dataset have been individually analyzed in the following subsections. As the ABC algorithm has some control parameters, their values were set and used in all experiments.

Result analysis for the Wisconsin Breast Cancer (WBC) dataset: The first experiment was performed on the Wisconsin Breast Cancer (WBC) dataset which is commonly used for deviation detection within the data mining community. This dataset contains 444 Benign and 39 Malignant objects which are assumed to represent the rare objects in this dataset. The outlier factor has been assigned the value of 0.1 to identify the top n outliers in this dataset. Table 1 shows the number of objects detected as deviation objects (top ratio), the number of malignant objects (correctly identified as deviation objects) and the number of benign objects. The results indicate that the ABC deviation detection algorithm can detect outliers with a top ratio of 6.21 and 48.72% detection rate. However, the maximum detection rate percentage is 89.74% with 11.20% top ratio.

Table 1: The results of WBC dataset

Number of objects (top ratio) (%)	Number of Malignant objects (DR %)	Number of Benign objects (FR %)
30 (6.21)	19 (48.72)	11 (2.48)
31 (6.41)	19 (48.72)	12 (2.70)
42 (8.70)	32 (82.05)	10 (2.25)
43 (11.20)	36 (89.74)	7 (1.58)

Table 2: The results of LYMP dataset

Number of objects (top ratio) (%)	Number of rare objects (DR %)	Number of normal objects (FR %)
5 (3.38)	2 (33.33)	3 (2.11)
6 (4.06)	3 (50.00)	3 (50.00)
7 (4.73)	3 (50.00)	4 (2.11)
6 (4.06)	4 (66.67)	2 (1.41)
12 (8.11)	6 (100.00)	6 (4.23)

Results analysis for the lymphography dataset: The second dataset used in the experiments is the lymphography dataset. This dataset was also used in many deviation detection studies in the literature. This dataset has 148 objects. The percentage of rare objects in this dataset is 4.24%. Accordingly, this dataset has 6 rare objects assumed to be deviation objects. The outlier factor has been assigned the value of 0.2 in this dataset. Table 2 shows the results obtained using ABC deviation detection on this dataset. The second and third columns of this table show the number of rare and normal records, respectively. Deviations are detected between 3.38 and 8.11% which are at the large percentage of the top n ratio. The algorithm could detect all outliers with a top ratio of 8.11 and 100% detection rate.

Result analysis for the Glass (GLS) dataset: The glass dataset is the last dataset used to evaluate the proposed method. This dataset has 9 objects assumed to belong to the rare objects. Table 3 shows a summarized evaluation based on detection rate and false alarm rate. The algorithm could detect 8 of the total outliers with a top ratio of 5.14%. This result means that the highest detection rate percentage was 88.89% with a 1.47% false alarm rate.

Comparative study: The algorithm has been run on three real datasets to illustrate its efficiency against two other methods. This section compares the proposed method (ABC deviation detection) and two other deviation detection methods: FindCBLOF by He *et al.* (2003) and TOD by Jiang *et al.* (2010). These methods are also Clustering-Based Deviation Detection Methods with two stages and use outlier factors of every object to identify which object is an outlier. All methods were compared using the same datasets (Wisconsin breast cancer and lymphography). The results obtained from the experiments were shown in tabular format for comparison with the other two methods. In comparing the three methods performances, the obtained results were analyzed based on deviation detection from the detection rate and the false alarm rate as in Table 4-7. In every search using the top n ratio or top n records, deviations are determined based on the detection rate. The detection rate measures how fast each method detects all outliers where the coverage ratio is 100%. The fast deviation detection speed represents a lower detection rate whereas a slow deviation detection speed represents a higher detection rate. Table 4 and 5 compare the proposed method with the FindCBLOF and TOD methods, respectively on the WBC dataset. The tables show that the FindCBLOF and TOD methods performances are slightly better than the proposed method. The FindCBLOF Method could detect all outliers with a top ratio of 13.25% and 64 objects and the TOD Method detected 38 deviations with a top ratio of 10.35% and 50 objects the ABC deviation detection algorithm could detect 36 of 39 objects with a top ratio of 11.20%. The

Table 3: The results of GLS dataset

Number of objects (top ratio) (%)	Number of rare objects (DR %)	Number of normal objects (FR %)
8 (3.74)	1 (11.11)	7 (3.43)
14 (6.54)	3 (33.33)	11 (5.39)
15 (7.00)	6 (66.67)	9 (4.41)
11 (5.14)	8 (88.89)	3 (1.47)

Table 4: The detection results on WBC dataset (ABC deviation detection vs. FindCBLOF)

ABC_deviation detection			FindCBLOF		
Number of objects (top ratio) (%)	Number of Malignant objects (%)	Number of Benign objects (%)	Number of objects (top ratio) (%)	Number of Malignant objects (%)	Number of Benign objects (%)
30 (6.21)	19 (48.72)	11 (2.48)	32 (6.63)	27 (69.23)	5 (1.13)
31 (6.41)	19 (48.72)	12 (2.70)	48 (9.94)	35 (89.74)	13 (2.93)
42 (8.70)	32 (82.05)	10 (2.25)	56 (11.59)	38 (97.44)	18 (4.05)
43 (11.20)	36 (89.74)	7 (1.58)	64 (13.25)	39 (100.00)	25 (5.63)

Table 5: The detection results on WBC dataset (ABC deviation detection vs. TOD)

ABC_deviation detection			TOD		
Number of objects (top ratio) (%)	Number of Malignant objects (%)	Number of Benign objects (%)	Number of records (top ratio) (%)	Number of Malignant objects (%)	Number of Benign objects (%)
30 (6.21)	19 (48.72)	11 (2.48)	33 (6.83)	31 (79.49)	2 (0.45)
31 (6.41)	19 (48.72)	12 (2.70)	40 (8.28)	35 (89.74)	5 (1.13)
42 (8.70)	32 (82.05)	10 (2.25)	44 (9.11)	36 (92.31)	8 (1.80)
43 (11.20)	36 (89.74)	7 (1.58)	50 (10.35)	38 (97.44)	12 (2.70)

Table 6: The detection results on LYMP dataset (ABC_deviation detection vs. FindCBLOF)

ABC_deviation detection			FindCBLOF		
Number of objects (top ratio) (%)	Number of rare objects (%)	Number of normal objects (%)	Number of objects (top ratio) (%)	Number of rare objects (%)	Number of normal objects (%)
5 (3.38)	2 (33.33)	3 (2.11)	7 (4.73)	4 (67)	3 (2.11)
6 (4.06)	3 (50.00)	3 (50.00)	22 (14.86)	4 (67)	18 (12.68)
7 (4.73)	3 (50.00)	4 (2.11)	30 (20.27)	6 (100)	24 (16.90)
10 (6.76)	6 (100.00)	4 (2.82)	-	-	-
12 (8.11)	6 (100.00)	6 (4.23)	-	-	-

Table 7: The detection results on LYMP dataset (ABC deviation detection vs. TOD)

ABC_deviation detection			TOD		
Number of objects (top ratio) (%)	Number of rare objects (%)	Number of normal objects (%)	Number of objects (top ratio) (%)	Number of rare objects (%)	Number of normal objects (%)
5 (3.38)	2 (33.33)	3 (2.11)	7 (4.73)	5 (83.33)	2 (1.41)
6 (4.06)	3 (50.00)	3 (50.00)	10 (6.76)	6 (100.00)	4 (2.82)
7 (4.73)	3 (50.00)	4 (2.11)	-	-	-
10 (6.76)	6 (100.00)	4 (2.82)	-	-	-
12 (8.11)	6 (100.00)	6 (4.23)	-	-	-

results indicate that the ABC_deviation detection algorithm presents a lower detection rate with fast speed in deviation detection for the WBC dataset than the other compared methods using the DR and FR measurements.

Table 6 and 7 show a summarized comparison between the proposed method and the two compared methods on the lymphography dataset. The comparison was also based on DR and FR. Table 6 shows that the proposed method detects the top n outliers with a top ratio of 6.76% whereas the FindCBLOF Method detects them with a top ratio of 20.27% which indicates that the proposed method is considerably faster than the FindCBLOF Method.

Moreover, the Table 7 shows a comparison between the proposed and TOD methods on the same dataset. The table shows that the proposed and TOD methods find the top n outliers with a top ratio of 6.76% which demonstrates the efficiency of the proposed method against that of the compared method.

Although, the FindCBLOF and TOD Methods gave higher detection rates than the ABC Deviation Detection Method, ABC deviation detection is faster in execution time.

CONCLUSION

Deviation detection is a data mining task with many application domains including the detection of fraud, e-commerce criminal activities, outbreak detection and computer network intrusion. The ABC algorithm was applied to solve the clustering problem and provided good results with regard to speed and converging to the best solutions. In this study, the ABC algorithm was reformulated to detect clustering-based deviations. The proposed method aims to cluster the dataset into many clusters detect data points that do not belong to any cluster and assign them as deviation objects. The performance of the Deviation Detection Method is measured with the Detection Rate (DR) and False Alarm Rate (FR) measurements. Two Cluster-based Deviation Detection methods are used for comparison with and evaluation of the proposed method. Experimentation on the proposed approach is performed on three benchmark datasets and demonstrated that the efficiency of the approach is better than those of other popular deviation detection techniques. The new algorithm achieved its performance with comparable results in top ratio and detection rate.

REFERENCES

Al-Zoubi, M.B., A. Al-Dahoud and A.A. Yahya, 2010. New outlier detection method based on fuzzy clustering. Wseas Trans. Inf. Sci. Applicat., 7: 681-690.

Alam, S., G. Dobbie, P. Riddle and M.A. Naeem, 2010. A swarm intelligence based clustering approach for outlier detection. Proceedings of the Congress on Evolutionary Computation, July 18-23, 2010, Barcelona, pp: 1-7.

Blake, C. and C. Merz, 2010. UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Ceglar, A., J.F. Roddick and D.M.W. Powers, 2007. CURIO: A fast outlier and outlier cluster detection algorithm for large datasets. Int. Workshop Integr. Artificial Intellig. Data Mining, 84: 37-45.

Ester, M., H.P. Kriegel, J. Sander and X. Xu, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, August 2-4, 1996, Portland, pp: 226-231.

Fathian, M., B. Amiri and A. Maroosi, 2007. Application of honey-bee mating optimization algorithm on clustering. Applied Math. Comput., 190: 1502-1513.

He, Z., X. Xu and S. Deng, 2003. Discovering cluster-based local outliers. Proceedings of the Annual Language Design and Implementation, June, 2003, San Diego, California.

Jiang, S., Q. Li, H. Wang and Y. Zhao, 2010. A two stage outlier detection method. Mini Micro Syst., 1: 1237-1240.

Jiang, S.Y. and A.M. Yang, 2009. Framework of clustering-based outlier detection. Proceedings of the 16th International Conference on Fuzzy Systems and Knowledge Discovery and Applications, August 16-18, 2009, Boston, MA.

Jiang, S.Y. and Q.B. An, 2008. Clustering-based outlier detection method. Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, October 18-20, 2008, Jinan, Shandong, China.

Karaboga, D. and B. Akay, 2009. A comparative study of artificial bee colony algorithm. Applied Math. Comput., 214: 108-132.

Karaboga, D. and B. Basturk, 2008. On the performance of Artificial Bee Colony (ABC) algorithm. Appl. Soft Comput., 8: 687-697.

Karaboga, D., 2005. An idea based on honey bee swarm for numerical optimization. Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, Kayseri/Turkiye. http://mf.erciyes.edu.tr/abc/pub/tr06_2005.pdf.

Laan, M.J., K.S. van der Pollard and J. Bryan, 2003. A new partitioning around medoids algorithms. J. Statist. Comput. Simulat., 73: 575-584.

Mansur, M.O. and M.N. Sap, 2005. Outlier detection technique in data mining: A research perspective. Postgr. Ann. Res., 1: 1-13.

Mohemmed, A.W., M. Zhang and W.N. Browne, 2010. Particle swarm optimisation for outlier detection. Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation Portland, OR, USA., July 7-11, 2010, ACM New York, USA., pp: 83-84.

Ng, R. and J. Han, 1994. Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, San Francisco, CA., USA., pp: 144-155.

Seidl, T., E. Muller, I. Assent and U. Steinhausen, 2009. Outlier detection and ranking based on subspace clustering. Proceedings of the Advances in Spatial and Temporal Databases, July 8-10, 2009, Aalborg, Denmark.

Selim, S.Z. and M.A. Ismail, 1984. K-means type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Mach. Intellg., 6: 81-87.

Tiwari, K., K. Mehta, N. Jain, R. Tiwari and G. Kanda, 2007. Selecting the appropriate outlier treatment for common industry applications. Proceedings of the Statistics and Data Analysis, April 16-19, 2007, Orlando, Florida.

Zhang, C., D. Ouyang and J. Ning, 2010. An artificial bee colony approach for clustering. Exp. Syst. Appl., 37: 4761-4767.

Zhang, J. and H. Wang, 2006. Detecting outlying subspaces for high-dimensional data: The new task, algorithms and performance. Knowledge Inf. Syst., 10: 333-355.

Zhang, T., R. Ramakrishnan and M. Livny, 1996. BIRCH: An efficient data clustering method for very large data bases. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 4-6, 1996, Cananda, pp: 103-114.

Zou, W., Y. Zhu, H. Chen and X. Sui, 2010. A clustering approach using cooperative artificial bee colony algorithm. Disc. Dyn. Nature Soc., http://www.hindawi.com/journals/ddns/2010/459796/.