http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (2): 338-344, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.338.344
© 2013 Asian Network for Scientific Information

A Novel Detector Generation Scheme for Detecting the Level of Abnormality of Equipment

Yinghui Liu, Shulin Liu and Yuzhen Li
School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China

Abstract: This study presented a novel scheme of detector generation and the concept of hyperring detector was proposed. Especially, the level of abnormality in equipment was detected with this new detector generation scheme. The reverse *k*- nearest neighbor algorithm and the *k*-means clustering algorithm were used in detector generation process. The presented method was experimented with both famous benchmark Fish's Iris data and real-world datasets. Preliminary results demonstrated that the new detector generation scheme had efficiency in detecting the level of abnormality in rolling bearing faults.

Key words: Abnormality detection, level of abnormality, detector generation scheme, hyper-ring detector, negative selection algorithm

INTRODUCTION

Early detection of abnormality of equipment is very crucial for its security and reliability. If abnormality level of equipment could be detected relatively accurate, the equipment running conditions will be known better. And it is essential in safety production and which could avoids major accidents (Shulin *et al*., 2002).

Since, Artificial Immune Systems (AIS) have been applied in the field of fault diagnosis, it has received a great deal of attention. Many mechanisms are applied in fault diagnosis and abnormality detection (Liu *et al*., 2012; Rasheed *et al*., 2012). The main character of AIS is that only normal data are needed to defense abnormalities. In 1994, the main character was named Negative Selection Algorithms (NSA) by Forest *et al*. (1994) and was firstly used in computer security and virus detection. From then on, NSA is widely used in anomaly detection in many areas. Shulin *et al*. (2002) firstly applied NSA in on-line fault diagnosis of rotary machinery and proved it is suitable for equipment fault diagnosis. Detectors originally expressed in binary form but fault diagnosis data are all represented in real-valued form. Dasgupta *et al*. (2004) investigated detectors in real-valued form which applied in man-in-the-loop aircraft operation fault detection. And the detectors in real-valued form are now widely used in NSA which applied in the field of fault diagnosis.

Though different variations of NSA have been frequently proposed, the main characteristics of this method described by Forest *et al*. (1994). These NSA variations are mostly concentrated on improving the algorithm performance via alternative detector generation schemes (Dasgupata *et al*., 2011).

In process of detectors representation method, the main purpose is using the least detectors but obtaining maximal non-self space coverage. After using real-valued form detectors, the detectors are firstly represented in hyper-spheres. Gao *et al*. (2006) introduced a detector optimization scheme. They used genetic algorithm to obtain the maximal possible radius of detector j without any overlapping with all the self samples. Ji and Dasgupata (2009) proposed detectors with variable-sized and variable-shaped which reduced false alarm rate as well as raised coverage in non-self space. Other researchers tried to represent detectors in hyper-ellipsoid (Shapiro *et al*., 2005) and in hyper-rectangular (Ostaszewski *et al*., 2006, 2007). Researchers have done a lot of outstanding works in detector representation method but there are still some problems. Firstly, it is still too many detectors for covering non-self space and the more detectors, the lower speed of algorithm. Secondly, for abnormality detection, detectors have no relationship with the level of abnormality. So most abnormality detection methods can only tell there is abnormal but can't tell how serious it is (Shulin *et al*., 2002).

If the border of self samples set could be recognized, the self space will be defined and the rest of the state space was considered to be abnormal space. The further the sample, the higher abnormality it will be. In this study, a novel detector generation scheme was introduced and the detectors were named hyper-ring detectors. The

Corresponding Author: Yinghui Liu, School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China

reverse *k*-nearest neighbor algorithm was used in recognizing the border of self samples. The *k*-means clustering algorithm was used in searching for the center of self samples set. Hyper-ring detectors were generated according to the center and border of self samples set. At the same time, the level of abnormality was calculated in accordance with hyper-ring detectors and the border of self samples set. The presented method was proved effective with famous benchmark Fish's Iris data and ball bearing test data from Case Western Reserve University (http://www.case.edu).

THE NOVEL DETECTOR GENERATION SCHEME

The reverse *k*-nearest neighbor algorithm: The Reverse *k*-Nearest Neighbor (RkNN) method was firstly proposed by Korn and Muthukrishnan (2000). It was effective in locating border of a database. In this study, the RkNN was used in recognizing the border of self samples set.

The RkNN algorithm was based on (kNN) algorithm and the distance between each two samples was calculated by Euclidian Distance. For N dimensional sample S_{id} ($S_{id} \in$ self samples set), its similarity with S_{jd} ($i \neq i$) is as follows:

$$\text{Sim}(i,j) = \frac{1}{\sqrt{\sum_{d=1}^{N}(S_{id} - S_{jd})^2}} \quad (1)$$

Self samples have relative low similarity were marked and its number is k_1. Then calculating reverse *k*-nearest neighbors for the k_1 self samples by RkNN and put the result in num. And judging whether the self sample is a boundary point according to threshold k_2. For example, if num$\leq k_2$, the self sample is a boundary, or it is not.

The RkNN in recognizing border of self samples was verified by two experiments in this study:

- The 1000 random distributed points in a circular area and 202 points on the border are used in testing. It was shown in Fig. 1

In order to recognize border of this circular area, parameter k_1 and k_2 should be chose properly. To choose parameter k_1 and k_2, a experiment that reflect relationship between different parameter combinations and the border recognition rate had been down. The result was shown in Table 1.

From Table 1, it could be seen that when $k_1 = 350$ and $k_2 = 270$, the border of circular area could be recognized relatively accurate. The recognition result was shown in Fig. 2:

Table 1: Choose different parameter combinations and the corresponding recognition rate in circular area (%)

	k_2			
k_1	202	230	250	270
250	79.20	84.15	89.11	91.09
300	85.36	91.08	94.05	96.53
350	94.05	94.55	97.52	100.00
400	81.23	91.58	96.53	100.00

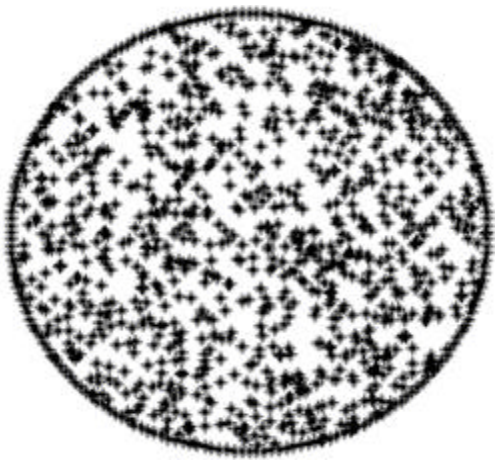

Fig. 1: Distributed map of random points in circular area

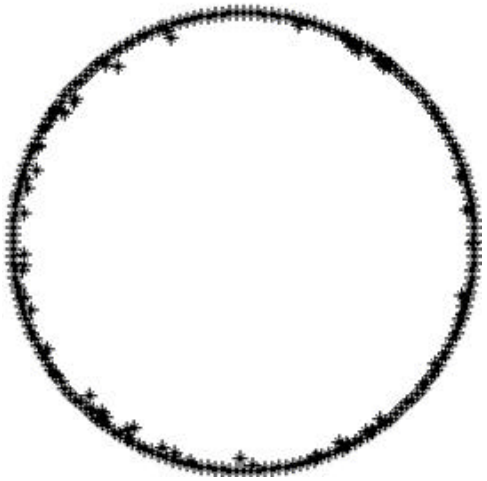

Fig. 2: Border of circular area when $k_1 = 350$ and $k_2 = 270$

- The 1000 random distributed points in a five-pointed star area and 280 points on the border are used in testing. It was shown in Fig. 3

Relationship between parameter combinations and border recognition rate in five-pointed area was shown in Table 2.

From Table 2, it could be seen that when $k_1 = 30$ and $k_2 = 400$, the recognition is pretty high. The border recognition result was shown in Fig. 4.

Fig. 3: Distributed map of random points in five-pointed star area

Fig. 4: Border of five-pointed area when k_1 = 30 and k_2 = 400

Table 2: Choose different parameter combinations and the corresponding recognition rate in five-pointed area (%)

	k_2			
k_1	280	310	360	400
10	81.89	88.53	93.32	95.43
20	84.55	90.29	94.22	96.51
30	89.53	91.44	96.41	99.18
50	73.08	81.46	87.69	95.78

From Table 1 and 2, it can be concluded that, for certain k_1, the recognition rate is raised with the addition of k_2 and for certain k_2, the recognition rate with k_1 increasing is firstly raise then decreased. In process of operation, the larger k_1 and k_2, the more time it needed. So, proper value of k_1 and k_2 have great effect on the performance of RkNN algorithm. According to the two experiments, RkNN was certified that it can be used in recognizing the border of self samples set.

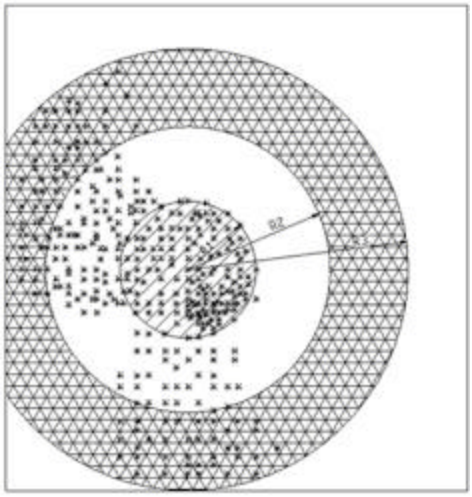

Fig. 5: Schematic diagram of a 2-level hyper-ring detector

***k*-means clustering algorithm:** Recent years, the *k*-means clustering algorithm was widely used in clustering (Xu *et al*., 2011; Garg and Jain, 2006; Tran *et al*., 2011; Jaradat *et al*., 2009), Image Compression (Venkateswaran and Rao, 2007; Mahi and Izabatene, 2011) and data preprocessing (Hemalatha and Vivekanandan, 2008). After recognizing border of self samples set, the *k*-means clustering algorithm was used in defining the center of self samples set in this study.

Detector generation method

Definition of hyper-ring detector: The center of self samples set C, interior radius R_t and outside radius R_{t+1} forms a t-level hyper-ring detector. T-level hyper-ring detector was represented as $d_{it} = [C, R_t, R_{t+1}]$. Figure 5 is shown the 2-level hyper-ring detector in two dimensions.

In Fig. 5, "*" represents self samples and slash area represents incomplete self space which is expressed as norm_space = $[C, R_1]$. Grid area represents 2-level hyper-ring detector which is expressed as $d_2 = [C, R_2, R_3]$. The area between incomplete self space and 2-level hyper-ring detector is 1-level hyper-ring detector, it is expressed as $d_1 = [C, R_1, R_2]$. Some self samples will be located in hyper-ring detector areas by this detector generation algorithm.

Process of hyper-ring detector generation: In this study, all of the samples are in N (NϵR) dimensional. The border of self samples set is $S_b = [S_{b1}; S_{b2};\cdots; S_{bn}]$ and the center of self samples set is $C = [c_1, c_2,\cdots, c_N]$. The detector generation steps are as follows:

Step 1: *k*-means clustering algorithm was used in calculating the center of self samples set

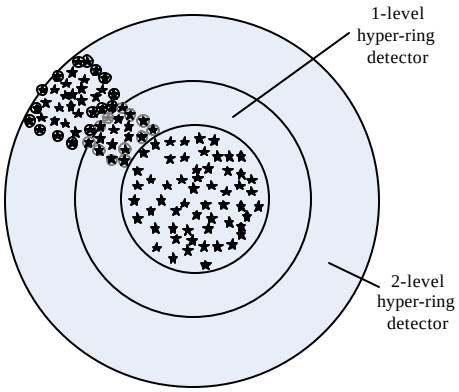

Fig. 6: Schematic diagram of hyper-ring detectors and self detectors

Step 2: RkNN was used in calculating the border of self samples set

Step 3: Euclidean Distance was used in calculating distance between any self samples on the border S_{bi} and the center C. The distance was:

$$dis\ (C,S_{bi}) = \sqrt{\sum_{j=1}^{N}(c_j - S_{bij})^2}\ \ i = 1,2,\cdots,n$$

and wrote down the distance in temp

Step 4: Find the minimum value min_temp in temp

Step 5: The interior diameter of 1-level hyper-ring detector is R_1 = min_temp+r_S and r_S is the radius of self samples

Step 6: Distance between center C and the furthest position in $[0, 1]^N$ is L, so step of each hyper-ring detector is:

$$dir = \frac{L - R_1}{t}$$

The t-level hyper-ring detector can be represented as $d_t = [C, ((t-1)\times dir+R_t), (t\times dir+R_t)]$.

It can be seen from Fig. 5, some self samples are located in hyper-ring detector area. So self samples have to be used in detecting abnormality and they were called self detectors. Definition of self detectors is as follows:

Step 1: For any self samples S_i, calculating distance between S_i and self samples center *C*. It was:

$$dis\ (S_i,C) = \sqrt{\sum_{j=1}^{N}(c_j - S_{ij})^2},$$

if $dis(S_i, C) \le R_1$, put these self samples in self_0. And if $dis(S_1, C) \in (((t-1)\times dir+R_1), (t\times dir+R_t)$, put them in self_t, (t = 1, 2,...)

Step 2: Calculating the border of self samples in self_t, t = 1, 2,... by RkNN algorithm. And put the results in $S'_{bt} = [S'_{bt1}; S'_{bt2}; \cdots; S'_{btp}]$, $(p = 1, 2, \cdots)$

Step 3: The border of self samples in t-lever (t = 1, 2,...) hyper-ring detector are considered to be self detectors, they were represented as $d_{ts} = [S'_{bt}, r_s, t]$

According to hyper-ring detectors and self detectors, computational time will be reduced. Schematic diagram of hyper-ring detectors and self detectors is shown in Fig. 6.

It can be seen form Fig. 6 that the self detectors were circled. They were circled by dark gray in 1-level hyper-ring detector area and they were circled by black in area of 2-level hyper-ring detector. The hyper-ring detectors and self detectors are used together to detect the level of abnormality in equipment.

SPECIFIC STEPS IN DETECTING ABNORMALITY

Proposed anomaly detection approach

Some definitions: Samples to be detected are $Ag_j = [ag_{j1}, ag_{j2},...ag_{jN})$. *C* is the center of self samples set. Similarity between hyper-ring detector and samples to be detected is as follows:

$$Similary\_1(ag_j, C) = \frac{1}{dis\ (ag_i, C) - R_1} \quad (2)$$

Where:

$$dis\ (ag_i, C) = \sqrt{\sum_{i=1}^{N}(ag_{ji} - c_i)^2}$$

For any self detector $d_{sk} = [S_{bk}, r_S, t]$ $S_{bk} \in$ self_t, t>0. Similarity between self detector and samples to be detected is:

$$Similary\_2\ (ag_j, d_s) = \frac{1}{\sqrt{\sum_{i=1}^{N}(ag_{ji} - s'_{bki})^2}} \quad (3)$$

The maximum similarity is max_t = max (Similary_2 (ag_j, D_t)). where, $D_t = [d_{S1}; d_{S2};...; d_{Sk}]$.

Intensity of anomaly and level of abnormality. First, judge samples to be detected in which hyper-ring detector area. And then calculate it's similarity to self detectors located in this hyper-ring detector area. For example, (1) if samples to be detected located in 1-level hyper-ring detector area and then calculating max_1. If:

$$max\_1 < \frac{1}{r_s},$$

intensity of anomaly is abnormality = dis (ag_i, C)-R_1 and the level of abnormality is 1, or intensity of anomaly is 0 and the level of abnormality is 0 too. (2) When samples to be detected located in t-level hyper-ring detector area, if:

$$max\_t < \frac{1}{r_s}$$

and similarity between self detectors located in lower than t-level hyper-ring detector area needed to be calculated. The maximum similarity between samples to be detected and self detectors within t-level hyper-ring detector area is max_t^{α}:

- When max_t^{α}<similary_1 (ag_j, C), intensity of anomaly is abnormality = dis (ag_i, C)-R_1 and the level of abnormality rank = t
- When $max_t^{\alpha} \geq$similarly_1(ag_j, C), the intensity of anomaly is:

$$abnomality = \frac{1}{max\_t^{\alpha}} - r_s$$

Under this situation, if:

$$\frac{1}{max\_t^{\alpha}} \in (0, R_1],$$

the level of abnormality rank = 1. And if:

$$\frac{1}{max\_t^{\alpha}} \in ((s-1)\times dir + R_1), (s * dir + R_1)] \ (s \geq 1),$$

the level of abnormality is rank = S+1

EXPERIMENTS AND RESULTS

Experiment of fisher iris data: The famous benchmark Fisher's iris data was used to illustrate and test the anomaly detection method. The iris data set contains 3 classes which named Setosa, Virginica and Versicolor separately. And there are 50 samples with 4 attributes in each class. Take the second and the third normalized attributes as horizontal ordinate and vertical ordinate. The iris data was shown in Fig. 7.

In Fig. 7, 40 samples of Setosa were used as self samples set represented by "○" and the rest 10 samples of Setosa, Virginica and Versicolor were used in testing △, ☆.

The level of abnormality was set to 3, r_S = 0.26 and dir = 0.5. Using RkNN in calculating the border of self samples set, parameters were set as k_1 = 10 and k_2 = 18. The recognition result of the intensity of anomaly is shown in Fig. 8.

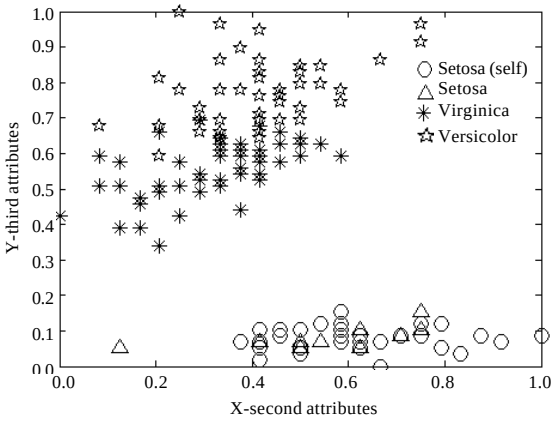

Fig. 7: Iris data shown in two dimensions

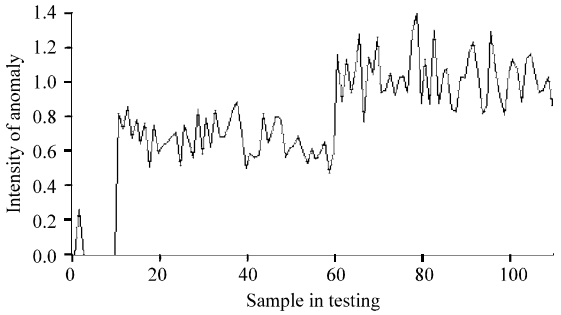

Fig. 8: The intensity of anomaly of Iris data

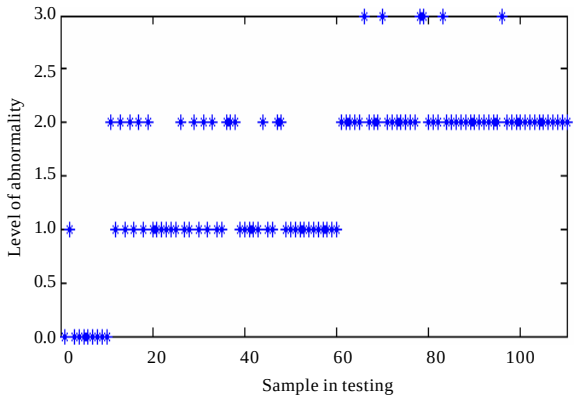

Fig. 9: The level of abnormality of Iris data

The level of abnormality is shown in Fig. 9.

From Fig. 8 and Fig. 9, it could be seen that one of the rest 10 samples of Setosa was assigned in 1-level anomaly. And because of it is a litter far away from self samples set which can be seen from Fig. 7. At last, the proposed detector generation scheme can be used in abnormality detection. In next section, the roller bearing data from Case Western Reserve University (http://www.case.edu) would be used to test performance of the method.

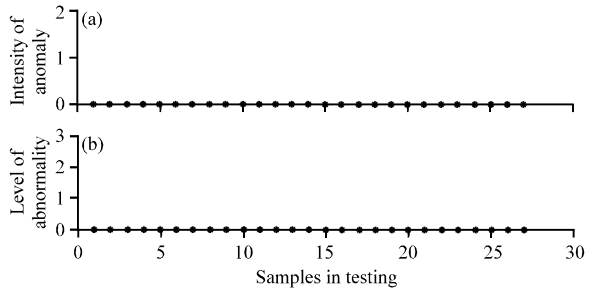

Fig. 10(a-b): The level of abnormality and intensity of anomaly of 28 self samples

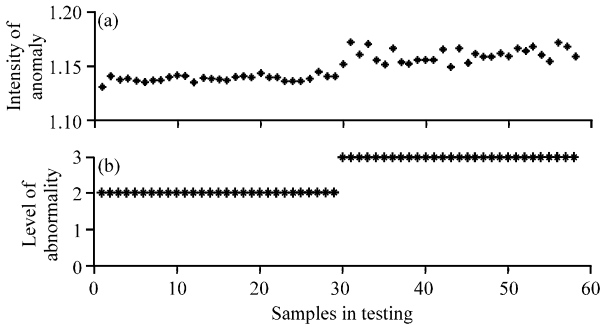

Fig. 11(a-b): The level of abnormality and intensity of anomaly of 58 inner race fault samples

Analysis of roller bearing: This study performs wavelet packet transform on the roller bearing signals to extract statistics of approximation coefficients that contain a great part of signal energy as sample-features. According to wavelet packet transform, 118 self samples were obtained, 29 inner race fault samples with 0.18 mm and 0.54 mm each and obtained 29 ball fault samples with 0.36 and 0.54 mm each.

Take 90 self samples randomly as training samples and all of the rest samples as testing samples. The level of abnormality is 4 and interior radius and outside radius for the first three hyper-ring detectors is (0.1,0.90),(0.90,1.30) and (1.30,1.50). There are three experiments.

Took the rest 28 self samples as testing samples, the testing result were shown in Fig. 10.

Took 58 inner race fault samples with 0.18 mm and 0.54 mm as testing samples, the testing result were shown in Fig. 11. Took 58 ball fault samples with 0.36 and 0.54 mm as testing samples, the testing result were shown in Fig. 12.

From Fig. 10, for self samples, they totally have no fault and their intensity of anomaly and the level of abnormality are all classified in 0. The deeper, the more serious fault the roller bearing will be. From Fig. 11, the intensity of anomaly and level of abnormality are all

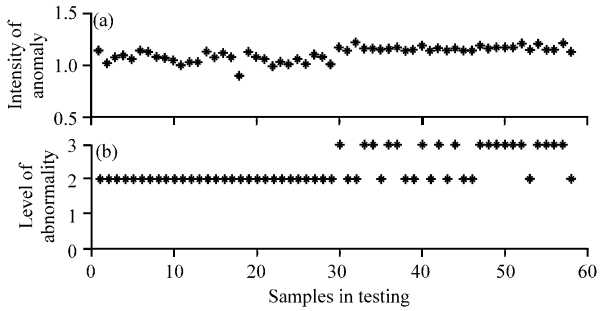

Fig. 12(a-b): The level of abnormality and intensity of anomaly of 58 bass fault samples

relatively low for the first 29 inner race fault samples with 0.18 mm. From Fig. 12, intensity of anomaly is relatively low for the first 29 ball fault samples with 0.36 mm and the level of abnormality can also reflect running condition. So if self samples are adequate and parameters are set properly, the level of abnormality detecting method in this study is efficient for abnormality detection. According to the level of abnormality, the running condition of roller bearing will be known perfectly.

CONCLUSION

Abnormality detection is important for state monitoring which is significant in production safety. For years, Negative Selection Algorithm (NSA) is widely used in anomaly detection in fault diagnosis. Detector in NSA is a hotspot especially in two aspects: the form of detectors and conflictions between the number of detectors and real-time property problem. In this study, a novel detector generation scheme was proposed based on the RkNN and *k*-means clustering method. According to the new scheme, the number of detectors could be reduced and could be perfectly used in abnormality detection. When there are enough self samples, this anomaly detection method could have good performance in defining the level of abnormality and which can tell us how the state of equipments was. For future work, more experiment will be down in actual equipment and make further validation of this abnormality detection method.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (51175316), the Specialized Research Fund for the Doctoral Program of Higher Education (20103108110006), Fundamental Research for key programs from Shanghai Committee of Science (11jc1404100) and Shanghai Talent Development Fund (047).

REFERENCES

Dasgupata, D., S.H. Yu and L.F. Nino, 2011. Recent advances in artificial immune systems: Models and applications. Applied Soft Comput., 11: 1574-1587.

Dasgupta, D., K. KrishnaKumar, D. Wong and M. Berry, 2004. Negative selection algorithm for aircraft fault detection. Artific. Immune Syst., 3239: 1-13.

Forest, S., S. Hofmeyr, A. Somayaji and T. Longstaff, 1994. Self-nonself discrimination in a computer. Proceedings of the Symposium on Computer Security and Privacy, May 16-18, 1994, IEEE Computer Society, USA., pp: 202-202.

Gao, X.Z., S.J. Ovaska and X. Wang, 2006. Genetic algorithms-based detector generation in negative selection algorithm. Proceedings of the IEEE Mountain Workshop on Adaptive and Learning Systems, July 24-26, 2006, Logan, UT, pp: 133-137.

Garg, S. and R.C. Jain, 2006. Variations of *k*-mean algorithm: A study for high-dimensional large data sets. Inform. Technol. J., 5: 1132-1135.

Hemalatha, M. and K. Vivekanandan, 2008. A semaphore based multiprocessing *k*-mean algorithm for massive biological data. Asian J. Sci. Res., 1: 444-450.

Jaradat, A., R. Salleh and A. Abid, 2009. Imitating *k*-means to enhance data selection. J. Applied Sci., 9: 3569-3574.

Ji, Z. and D. Dasgupata, 2009. V-Detector: An efficient negative selection algorithm with "Probably Adequate" detector coverage. Inform. Sci., 179: 1390-1406.

Korn, F. and S. Muthukrishnan, 2000. Influence sets based on reverse nearest neighbor queries. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 201-212.

Liu, S., Y. Liu, Y. Tang and R. Jiang, 2012. A novel pattern recognition approach based on immunology. Inform. Technol. J., 11: 134-140.

Mahi, H. and H.F. Izabatene, 2011. Segmentation of satellite imagery using RBF neural network and genetic algorithm. Asian J. Applied Sci., 4: 186-194.

Ostaszewski, M., F. Seredynski and P. Bouvry, 2006. Immune anomaly detection enhanced with evolutionary paradigms. Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, July 8-12, 2006, Seattle, Washington, USA, pp: 119-126.

Ostaszewski, M., F. Seredynski and P. Bouvry, 2007. Coevolutionary-based mechanisms for network anomaly detection. J. Math. Model. Algor., 6: 411-431.

Rasheed, M.M., O. Ghazali and N.M. Norwawi, 2012. Intelligent signature detection for scanning internet worms. Inform. Technol. J., 11: 760-767.

Shapiro, J.M., G.B. Lamont and G.L. Peterson, 2005. An evolutionary algorithm to generate hyper-ellipsoid detectors for negative selection. Proceedings of the Conference on Genetic and Evolutionary Computation, June 25-29, 2005, Washington, DC, USA., pp: 337-344.

Shulin, L., Z. Jiazhong, S. Wengang and H. Wenhu, 2002. Negative-selection algorithm based approach for fault diagnosis of rotary machinery. Proceedings of American Control Conference, Vol. 5, May 8-10, 2002, The Netherlands, pp: 3955-3960.

Tran, T.T., H.M. Cho and S.B. Cho, 2011. A robust method for detecting lane boundary in challenging scenes. Inform. Technol. J., 10: 2300-2307.

Venkateswaran, N. and Y.V.R. Rao, 2007. *k*-means clustering based image compression in wavelet domain. Inform. Technol. J., 6: 148-153.

Xu, W., Z. Qin and Y. Chang, 2011. A framework for classifying uncertain and evolving data streams. Inform. Technol. J., 10: 1926-1933.