EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Statistical wavelet-based anomaly detection in big data with compressive sensing

Wei Wang*, Dunqiang Lu, Xin Zhou, Baoju Zhang* and Jiasong Mu

Abstract

Anomaly detection in big data is a key problem in the big data analytics domain. In this paper, the definitions of anomaly detection and big data were presented. Due to the sampling and storage burden and the inadequacy of privacy protection of anomaly detection based on uncompressed data, compressive sensing theory was introduced and used in the anomaly detection algorithm. The anomaly detection criterion based on wavelet packet transform and statistic process control theory was deduced. The proposed anomaly detection technique was used for through-wall human detection to demonstrate the effectiveness. The experiments for detecting humans behind a brick wall and gypsum based on ultra-wideband radar signal were carried out. The results showed that the proposed anomaly detection algorithm could effectively detect the existence of a human being through compressed signals and uncompressed data.

Keywords: Anomaly detection; Big data; Through-wall human detection; Compressive sensing

1 Introduction

Anomaly detection refers to finding inconsistency with the desired pattern in data, which is also known as novelty detection, anomaly mining, and noising mining. Anomaly detection has many applications, such as credit card fraud detection, medical diagnostics information anomaly detection, industrial equipment fault detection and structural defect detection, network intrusion detection, and a novel theme of text mining. Currently, anomaly detection methods include anomaly detection based on classification, anomaly detection based on the nearest neighbor method, anomaly detection based on clustering, statistical anomaly detection, anomaly detection based on information theory, spectral theory anomaly detection, and so on [1]. All the anomaly detection techniques can operate in one of the three modes: supervised anomaly detection which assumed the availability of a training data set that labeled instances for normal as well as anomaly class, semi-supervised anomaly detection which only assumed the availability of a normal data set, and unsupervised anomaly detection which did not require training data.

'Big data' refers to large, diverse, complex, longitudinal, and distributed data sets generated from instruments, sensors, internet transactions, email, video, click streams, and other digital sources available today and in the future. Some core technologies are needed to solve the problem in big data, such as effective data collection and storage and data mining techniques and anomaly detection method. However, the existing anomaly detection algorithms are mainly based on the complete data. It greatly limits the application of anomaly detection algorithms in big data [2,3]. The main reason is that big data acquisition and storage become increasingly difficult with the increasing amount of data because of the sampling bandwidth and storage space constraints. Meanwhile, anomaly detection based on the complete data would inevitably expose some sensitive data which the data owner did not want to be leaked, so the effective privacy protection is the problem needed to be overcome currently.

Compressive sensing (or compressed sampling, CS) theory suggested that a high-dimensional signal can be projected into a low-dimensional space with a random measurement matrix when the signal was sparse or compressible which was proposed by Donoho and Candès in 2006 [4,5]. Then, the original signal can be reconstructed from the low-dimensional information by solving an optimization problem. In other words, the low-dimensional signal contained the main features of the original signal, so

* Correspondence: wangweivip@tju.edu.cn; wdxyzbj@163.com
College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China

CS theory can provide an effective method for anomaly detection in big data.

Usually, the number of normal data instances is more than the number of anomaly data instances. This just meets the sparsity requirements of compressive sensing theory. In [6], a compressed sensing-based clone identification scheme for sensor networks has been proposed according to the sparsity of clone appearance in the network. The data was recovered at the data base station, and then the safety of sensor networks was monitored. The algorithm achieved the lowest communication overhead. In [7], the compressed sensing theory was used for wireless network data acquisition and achieved the anomaly detection during a signal recovery procedure based on the modified BP reconstruction algorithm. Moreover, considering the asymmetry of anomaly and energy consumption, an improvement was made by giving a distributed detection algorithm. However, the main purpose of compressive sensing theory in the above research was to reduce the pressure of data storage and transmission, but it did not realize the anomaly detection in the compressed domain and cannot overcome the large amount of data calculation and privacy protection issues. Saha used the compressed sensor network data for anomaly detection based on the spectrum theory method and obtained satisfactory detection results in the light of residual analysis of compressed data [8,9]. In [10,11], random projection in conjunction with principal component analysis (PCA) was implemented for anomaly detection in the compressed domain. An application of this proposed methodology to detect IP-level volume anomalies in the computer network traffic suggested a high relevance to practical problems. Although this type of methodologies can achieve the anomaly detection in the compressed domain and has privacy protection features, the algorithms were complex, computationally intensive, and insufficient in real-time detection.

According to the problem in anomaly detection of big data, an anomaly detection algorithm in the compressed domain is proposed which makes full use of the advantages of compressive sensing technology. The experiment for through-wall human detection with the provided algorithm is carried out to test the effectiveness of the algorithm. The remainder of the paper is organized as follows. In Section 2, the criterion and procedure of anomaly detection are deduced base on wavelet packet transform and statistical method. The compressive sensing theory is then introduced. Experimental results for the detection of a human being will be shown in Section 3. Conclusion and discussion are in Section 4.

2 Background and anomaly detection procedure

2.1 Anomaly detection procedure

The conventional signal frequency spectrum analysis is built on the basis of Fourier transform (FT) which is a global transform. However, FT has a low-frequency resolution and cannot recognize the subtle changes of the frequency spectrum. Wavelet transform (WT) may be viewed as an extension of the traditional FT with adjustable window locations and sizes. Compared with the Fourier-based analyses that use global sine and cosine functions as bases, the basis wavelets are local functions, each of which is defined by two parameters: its scale (relating to frequency) and its position (relating to time). One possible drawback of the WT is that its frequency resolution is quite poor in the high-frequency region. The wavelet packet transform (WPT) is one extension of the WT that provides a complete level-by-level decomposition. The WPT enables the extraction of features from signals that combine stationary and nonstationary characteristics with an arbitrary time-frequency resolution.

The wavelet packet transform of a time-domain signal $x(t)$ can be calculated using a recursive filter-decimation operation. After the signal $x(t)$ is decomposed into j levels of decomposition and the node signals are reconstructed as $x_j^i(t)$, the signal $x(t)$ can be expressed as

$$x(t) = \sum_{i=1}^{2^j} x_j^i(t) \tag{1}$$

The node signal energies E_j^i can be defined as

$$E_j^i = \int_{-\infty}^{\infty} x_j^i(t)^2 dt = \Sigma x_j^i(t)^2 \tag{2}$$

According to the theory of WPT, each node signal contains information of the original signal in a specific time-frequency window. Hence Equation 2 illustrates that the node signal energy E_j^i is the energy stored in the corresponding frequency band. Obviously, the frequency components will be varied when anomaly occurs in the original signal. Thus, the anomaly detection can be achieved by investigating the changing trend of E_j^i.

The wavelet packet components with small energy magnitudes are easily jammed by the measurement noise. Thus, in this paper, instead of directly observing node signal energies on an individual basis, the criterion for anomaly detection is designed as

$$\mathrm{ADC} = \sum_{i=1}^{m} \frac{|\Delta E_j^i - \overline{\Delta E_j^i}|}{\overline{\Delta E_j^i}} \tag{3}$$

where ΔE_j^i is the node signal energy ratio in the total signal energy and $\overline{\Delta E_j^i}$ is the reference baseline of node signal energy ratio which is the mean value of node signal energy by measuring some subsequent signals. In order to eliminate the noise effect, the first m dominant nodes are retained. It can be deduced that anomalies in the signal would affect the wavelet node signal energies and subsequently alter the

criterion 'ADC.' However, the anomalies are not the only factor that can affect the criterion which also can be influenced by measurement noise. It is essential to establish threshold values for the anomaly detection criterion so that the criterion can be used to extract anomalies from measurement noise with a large probability [11]. In this paper, the threshold values would be fixed based on statistical process control (SPC) and statistical process control charts which are used to describe the output characteristics of the process in a coordinate graph.

SPC was proposed by Dr. W.A. Shewhart, and the statistical process control charts described the output characteristics of the process in a coordinate graph. The algorithm was proposed to monitor the manufacturing process so as to reduce variation and guarantee quality into the product. The control chart includes center line, upper control limit (UCL), and lower control limit (LCL). For anomaly detection applications, the core of the technique is to establish the control limits that can enclose variation of the extracted criterion due to measurement noises with a large probability.

Assume continuous measuring of m sets of time-domain signals under the same status. In other words, there are no anomalies during the measurements. According to Equations 1, 2, and 3, a total of p ADCs can be acquired using the average node energies as the reference baseline. Furthermore, the mean values and the standard deviation of p ADCs can be obtained as μ_{ADC} and S_{ADC}. On the basis of SPC theory, an X-bar control chart is used to determine threshold values of ADC. Suppose that the p ADCs are divided into subgroups of size q. Then, the one-side $1-\alpha$ upper confidence level for average ADCs of a subgroup can be defined as

$$\mathrm{UCL}_{\alpha} = \mu_{\mathrm{ADC}} + Z_{\alpha}\left(\frac{S_{\mathrm{ADC}}}{\sqrt{q}}\right) \tag{4}$$

where Z_{α} is the value of a standard normal distribution with zero mean and unit variance such that the cumulative probability is 100 $(1-\alpha)$%. The level UCL_{α} can be regarded as the threshold value of the criterion. Therefore, if no anomalies happen, the average criterion ADC of a followed subgroup would be in the range of UCL_{α} with a high probability of 100 $(1-\alpha)$%. On the other hand, when the average criterion ADC of a consecutive subgroup is beyond the limit, it shows that there are some anomalies. However, it should indicate that the SPC is a statistical principle of hypothesis testing. So, there are two types of hypothesis testing errors. Usually, the confidence limit can be improved by increasing the sizes p and q. From the anomaly detection procedure, it can be seen that no training data is required to construct a mathematical mode for anomaly detection. That is to say that the proposed algorithm belongs to unsupervised anomaly detection and can achieve online anomaly detection.

2.2 Compressive sensing theory

Although the proposed algorithm can overcome the noise interference and achieve anomaly detection with high probability, this algorithm is based on the complete data which greatly limits the application in the big data field. Compressive sensing theory overwhelms the limitation of Nyquist sampling theory and can acquire and compress data simultaneously. The theory provides a feasible basis for the proposed anomaly detection algorithm in the big data field.

For signal $x \in R^N$, it can be expressed as

$$x = \sum_{i=1}^{N} \phi_i \theta_i \quad \text{or} \quad x = \Phi\theta \tag{5}$$

where Φ is the $N \times N$ orthonormal transform basis and θ is the expansion coefficient vector under the orthonormal basis. If signal x is a K sparse signal, that is, K elements in

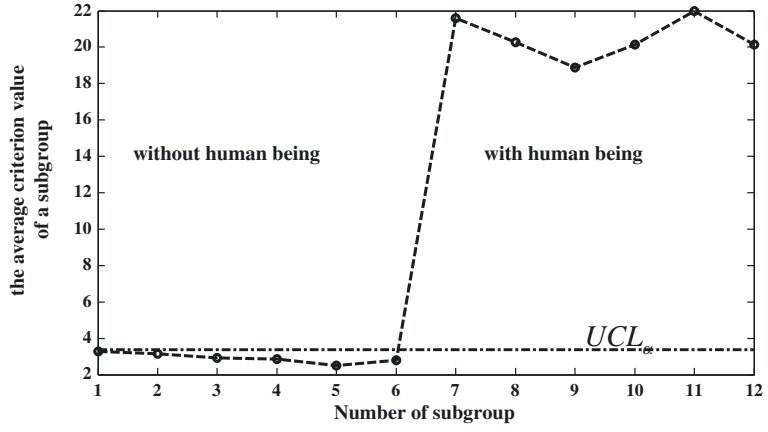

Figure 1 Through-wall human detection with complete data for brick wall.

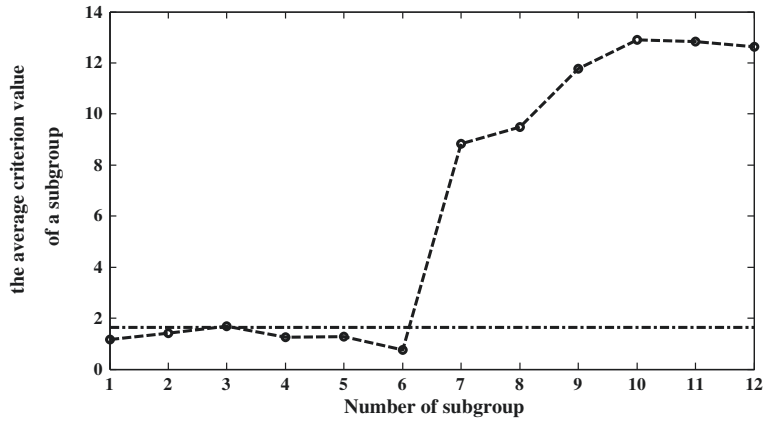

Figure 2 Through-wall human detection with complete data for gypsum wall.

vector θ are not zero and K is far less than N, the signal x can be collected with a small set of nonadaptive, linear measurements according to compressive sensing theory. Then, it can be described as follows [12-18]:

$$y = \Psi x = \Psi\Phi\theta \tag{6}$$

where Ψ is a $M \times N$ random measurement matrix and $M < N$. Here, (Φ, Ψ) is a pair of orthobases which followed the incoherence restriction.

When the above condition holds, the expansion coefficients θ can be reconstructed by solving the following l_0-norm constrained optimization problem:

$$\widehat{\theta} = \arg\min\|\theta\|_0 \quad \text{s.t.} \quad y = \Phi\Psi\theta \tag{7}$$

where the $\|\theta\|_0$ norm counts the number of nonzero components of θ. However, solving Equation 7 was both numerically unstable and NP complete. Instead of solving the l_0 minimization problem, the nonadaptive CS theory seeks to solve the 'closest possible' tractable minimization problem, i.e., the l_1 minimization:

$$\widehat{\theta} = \arg\min\|\theta\|_1 \ \text{s.t.} \ y = \Phi\Psi\theta \tag{8}$$

If x is termed as K sparse in the orthonormal basis, then we only need to collect $M = O(K \log(N/K))$ random measurements to recover the signal by l_1-norm algorithm. Many recovery algorithms based on linear programming, convex relaxation, and greedy strategies have been proposed to solve Equation 8, such as matching pursuit (MP), orthogonal matching pursuit (OMP), StOMP, subspace pursuit (SP), and CoSaMP. Finally, the reconstruction signal x can be given by $\hat{x} = \Phi\hat{\theta}$.

According to compressive sensing theory, the acquired low-dimensional signal contained the main features of the signal under the premise of appropriate measurement matrix. So, the frequency component features would be included in the gathered low-dimensional signal. Therefore, the proposed anomaly detection procedure for big data could be carried out in the compressed domain.

3 Experiment results

The performance of our approach was evaluated by using experiments of through-wall human detection.

3.1 Through-wall human detection based on UWB radar

Through-wall human detection is of great interest for many applications, such as military reconnaissance, anti-terrorism, and medical and natural disaster ambulance, which can penetrate non-metallic media to detect life signal in far-off areas. Due to the strong anti-interference ability, high-resolution performance, and good target recognition capabilities, the ultra-wideband (UWB) radar has emerged as one of the most optimal choices for through-wall human detection which can emit very short duration pulses to penetrate walls, bulkhead, and other obstacles. The through-wall detection of a human being is based on

Table 1 Through-wall human detection with compressed data for gypsum wall

		Average criterion value of a subgroup					
Compressive ratio 1:5 $UCL_\alpha = 0.973$	Without a human being	0.9231	0.7760	0.6595	0.6153	0.5798	0.6996
	With a human being	128.205	128.433	133.584	134.44	133.735	131.339
Compressive ratio 1:3 $UCL_\alpha = 1.289$	Without a human being	1.256	0.9657	0.7574	0.7075	0.9152	0.9624
	With a human being	23.62	24.15	24.06	23.63	23.46	23.47

Table 2 Through-wall human detection with compressed data for brick wall

		Average criterion value of a subgroup					
Compressive ratio 1:5 $UCL_\alpha = 1.1419$	Without a human being	1.1109	0.9593	0.8095	0.8438	0.9436	0.7276
	With a human being	42.7864	43.0531	44.7161	45.0177	44.9143	44.9054
Compressive ratio 1:3 $UCL_\alpha = 1.1938$	Without a human being	1.1035	1.0675	1.1122	1.0067	0.8548	0.9219
	With a human being	60.3670	61.4609	63.9775	64.5992	65.3513	64.9036

the fact that the human body is always in a state of motion even if it sleeps or is trapped because of breathing. These tiny human motions would cause the scattering and reflection changes of an electromagnetic wave which is emitted by the UWB radar and passes through walls to reach the human body target. Furthermore, human target detection, location, and tracking can be achieved through signal extraction and analysis of human characteristics from the radar echo signal [19-21]. Qilian Liang from the University of Texas at Arlington has investigated these topics deeply [22-29].

However, the UWB echo signals usually are the big data which increase the burden of data sampling and storage, and it is difficult for anomaly detection or through-wall human detection with complete UWB echo signal data. In this paper, we would use the proposed anomaly detection procedure in the compression domain for through-wall human detection to verify its feasibility.

3.2 Experimental results and analysis

In the experiments, we used the P220 UWB radar of Time Domain Company as the detection tool, and it worked in monostatic mode wherein the waveform pulses were transmitted from a single omnidirectional antenna and scattered waveforms were acquired by a collocated omnidirectional antenna. In the project, the P220 UWB radar was working with a center frequency of 4.3 GHz. In this project, data were collected at three different locations with different types of walls which were brick wall and gypsum wall. The radar parameters were the same and were as follows: hardware integration = 512, software integration = 2, pulse repetition frequency 9.6 MHz, step size 13 bin, and window size 10 ft. The echo signals were collected at two statuses: no person behind the wall and a still human being standing behind the wall.

To verify the feasibility of the proposed algorithm, each of the 30 sets of echo signals having a status of without a human being and with a human being was acquired. The signals were decomposed by an eight-layer wavelet packet with a 'db10' wavelet. The first 20 node signals were used to construct the criterion ADC. The length of subgroup is $q = 5$. The upper confidence level is $1 - \alpha = 0.95$. The random Gaussian measurement matrix was chosen for anomaly detection in the compression domain based on compressive sensing.

The experimental results of through-wall human detection for complete data and compressed data with compression ratios of 1:3 and 1:5 were shown in Figures 1 and 2 and Tables 1 and 2.

According to the proposed anomaly detection procedure, from the experimental results, it can be seen that the identification method could effectively detect the existence of a human being in complete data and compressed data with a high probability of 95%. Meanwhile, the results showed that the difference of the average criterion of a subgroup between without a human being and with a human being became more obvious in the compressed domain. It is mainly because more features of the original signal have been eliminated in the compressed domain and affected the reference baseline of the node signal energy $\bar{\Delta E_j^i}$. Otherwise, due to the random feature of the measurement matrix, we got the average criterion value of a subgroup by calculating three times and acquired the mean value as the criterion value.

4 Conclusions

Big data is a collection of data sets which are too large and complex to be processed using traditional data processing methods. Anomaly detection is an important problem that has been researched within diverse research areas and application domains. However, traditional anomaly detection techniques based on complete data are confronted with many difficulties for big data.

According to the principle that the frequency components would vary because of the existence of anomaly, the paper proposed an anomaly detection algorithm based on wavelet packet transform and statistic process control theory. Due to the sampling and storage burden of anomaly detection in big data, compressive sensing theory was used in the proposed anomaly detection method. Subsequently, we proved that through-wall human detection using compressed data could achieve equivalent performance as it did using the original uncompressed data and reduced the computational cost significantly based on the proposed anomaly detection technique.

Competing interests

The authors declare that they have no competing interests.

Acknowledgment
The authors would love to thank Professor Qilian Liang from the University of Texas at Arlington for providing the UWB radar data. This research was supported by the Tianjin Younger Natural Science Foundation (12JCQNJC00400) and National Natural Science Foundation of China (61271411).

Received: 25 September 2013 Accepted: 1 November 2013
Published: 17 November 2013

References

1. V Chandola, A Banerjee, V Kumar, Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–72 (2009)
2. CT Chou, R Rana, W Hu, Energy efficient information collection in wireless sensor networks using adaptive compressive sensing, in *IEEE 34th Conference on Local Computer Networks, Zurich, Switzerland, 20–23 October 2009*, pp. 443–450
3. J Wang, S Tang, B Yin et al., Data gathering in wireless sensor networks through intelligent compressive sensing, in *Proceedings IEEE INFOCOM, Orlando, FL, 25–30 March 2012*, pp. 603–611
4. DL Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
5. E Candès, Compressive sampling, in *Proceedings of International Congress of Mathematicians* (European Mathematical Society Publishing House, Madrid, 2006), pp. 1433–1452
6. CM Yu, CS Lu, SY Kuo, CSI: compressed sensing-based clone identification in sensor networks, in *8th IEEE International Workshop on Sensor Networks and Systems for Pervasive Computing 2012, Lugano, 9–23 March 2012*, pp. 296–301
7. Y Xia, Z Zhao, H Zhang, Distributed anomaly event detection in wireless networks using compressed sensing, in *Proc IEEE ISCIT 2011, Hangzhou, 12–14 October 2011*, pp. 250–255
8. S Budhaditya, DS Pham, M Lazarescu, S Venkatesh, Effective anomaly detection in sensor networks data streams, in *9th IEEE International Conference on Data Mining, Miami, FL, 6–9 December 2009*, pp. 722–727
9. DS Pham, S Venkatesh, M Lazarescu, S Budhadity, Anomaly detection in large-scale data stream networks. Data Min. Knowl. Discov. (2012). doi: 10.1007/s10618-012-0297-3
10. Q Ding, ED Kolaczyk, A compressed PCA subspace method for anomaly detection in high-dimensional data. IEEE Trans. Inf. Theory **59**(11), 7419–7422 (2012)
11. DS Pham, B Saha, M Lazarescu, S Venkatesh, *Scalable network-wide anomaly detection using compressed data, Technical report* (Institute for Multisensor Processing and Content Analysis, Department of Computing, Curtin University of Technology, 2009)
12. Z Sun, CC Chang, Statistical wavelet-based method for structural health monitoring. J. Struct. Eng. **130**(7), 1055–1062 (2004)
13. E Candès, M Wakin, An introduction to compressive sampling. IEEE Signal Process Mag. **25**, 21–30 (2008)
14. L Xu, Q Liang, X Cheng, D Chen, Compressive sensing in distributed radar sensor networks using pulse compression waveforms. EURASIP J. Wirel. Commun. Netw. (2013). doi: 10.1186/1687-1499-2013-36
15. L Xu, Q Liang, Zero correlation zone sequence pair sets for MIMO radar. IEEE Trans. Aerosp. Electron. Syst. **48**(3), 2100–2113 (2012)
16. L Xu, Q Liang, Orthogonal pulse compression codes for MIMO radar system, in *IEEE Globecom, Miami, FL, 6–10 Dec. 2010*
17. L Xu, Q Liang, Waveform design and optimization in radar sensor network, in *IEEE Globecom, Miami, FL, 6–10 Dec. 2010*
18. Q Liang, JM Mendel, Design interval type-2 fuzzy logic systems using SVD-QR method: rule reduction. Int. J. Intell. Syst. **15**(10), 939–957 (2000)
19. S Singh, Q Liang, D Chen, L Sheng, Sense through wall human detection using UWB radar. EURASIP J. Wirel. Commun. Netw. **2011**, 20 (2011)
20. D Kocur, J Rovňáková, M Švecová, Through wall tracking of moving targets by M-sequence UWB radar. Stud. Comput. Intell. **243**, 349–364 (2009)
21. W Wang, X Zhou, B Zhang, J Mu, Anomaly detection in big data from UWB radars. Secur. Comm. Network (2013). doi: 10.1002/sec.745
22. L Xu, Q Liang, Radar sensor network using a set of new ternary codes: theory and application. IEEE Sensors J. **11**(2), 439–450 (2011)
23. Q Liang, Situation understanding based on heterogeneous sensor networks and human-inspired favor weak fuzzy logic system. IEEE Syst. J. **5**(2), 156–163 (2011)
24. Q Liang, Biologically-inspired target recognition in radar sensor networks. EURASIP J. Wirel. Commun. Netw. **2010**, 523435 (2010)
25. Q Liang, X Cheng, S Samn, NEW: network-enabled electronic warfare for target recognition. IEEE Trans. Aerosp. Electron. Syst. **46**(22), 558–568 (2010)
26. Q Liang, X Cheng, KUPS: knowledge-based ubiquitous and persistent sensor networks for threat assessment. IEEE Trans. Aerosp. Electron. Syst **44**(3), 1–7 (2008). 1060–1069
27. Q Liang, Automatic target recognition using waveform diversity in radar sensor networks. Pattern Recogn. Lett. **29**(2), 377–381 (2008)
28. Q Liang, Radar sensor networks: algorithms for waveform design and diversity with application to ATP with delay-doppler uncertainty. EURASIP J. Wirel. Commun. Netw. 1–9 (2007)
29. Q Liang, Waveform design and diversity in radar sensor networks: theoretical analysis and application to automatic target recognition, in *IEEE Third Annual Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 2006), Reston, VA, 28–28 September 2006*

doi:10.1186/1687-1499-2013-269
Cite this article as: Wang *et al.*: **Statistical wavelet-based anomaly detection in big data with compressive sensing.** *EURASIP Journal on Wireless Communications and Networking* 2013 **2013**:269.